NOTE TO USERS

Page(s) not included in the original manuscript and are unavailable from the author or university. The manuscript was scanned as received.

125-128,133-139

This reproduction is the best copy available.

UMI[®]

**Evolutionary Genetics of Wood Formation in Plants: Sequence analysis
and Characterization of the Plant Ethylene Receptor and Subtilisin-like
Protease Protein Families**

By

Christopher Paul Kozela

Department of Biology

College of Natural Sciences

Carleton University

Ottawa, Ontario

**Dissertation submitted to Faculty of Graduate Studies, Carleton
University in partial fulfilment of the requirements for the degree of
Master of Science**

2003

© Copyright Christopher Paul Kozela 2003

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisitions et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-89021-X

Our file Notre référence

ISBN: 0-612-89021-X

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this dissertation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de ce manuscrit.

While these forms may be included in the document page count, their removal does not represent any loss of content from the dissertation.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

The undersigned hereby recommends to the
Faculty of Graduate Studies and Research
acceptance of the thesis

**Evolutionary Genetics of Wood Formation in Plants: Sequence analysis
and Characterization of the Plant Ethylene Receptor and Subtilisin-like
Protease Protein Families**

Submitted by
Christopher Paul Kozela

in partial fulfilment of the requirements
for the degree of Masters of Science

Chair, Department of Biology

Thesis Supervisor

Carleton University

Abstract

The generation of water-conducting tissues in plants is a highly regulated developmental process. Several candidate genes were identified as being potentially involved in xylem fiber development in *Arabidopsis*. An evolutionary analysis of two of these candidate genes was undertaken using molecular sequence data from various plant species. The plant ethylene receptors show highly ordered relationships when protein sequences were examined using a molecular phylogenetic approach. These cluster into 3 groups suggesting that at least 3 unique sets of functional constraints may exist. These groups correlate with physiological responses of different ethylene receptor mutants in *Arabidopsis* under environmental stress. The sequence relationships of 102 subtilisin-like proteases from various plant species were also examined. These proteins show an extremely diverse history based on protein sequence comparisons and have been grouped into 8 sequence groups. The biological significance of these groups is discussed and a structure for molecular sequence data visualization is presented.

Acknowledgements

First of all I would like to thank all my friends and family who have supported me over the last 28 plus years. Oh the strength you must have.

I would like to thank all of the students, staff, and faculty, past and present, in the Department of Biology for their encouragement and for the environment they produced during my stay here.

I would like to thank my thesis advisors Brian Miki and Linda Bonen for their encouragement and assistance in producing this work.

Peace to the Reganites, past and present.

Peace to the 21 Crew, good times.

Thanks to Byron Johnson, for your friendship, guidance, and encouragement over the years.

I especially thank my parents, Paul and Slava Kozela. You have always been there when I needed you and even though I may not always show it, I am grateful.

Finally I thank Sharon Regan. This could not have been possible without you Sharon. Your support of my curiosity over the past two years has made any difficulties seem easily worth it. I value your teaching, your support, and most of all your friendship. Thank you for everything.

Table of Contents

Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
List of Abbreviations	x
List of Appendices	xi
1. General Introduction	1
2. How Plants Make Tubes	3
2.1. Abstract	3
2.2. Introduction	4
2.3. Tracheary Element Formation	10
2.4. Lysigenous Aerenchyma Share Developmental Characteristics With TE's	14
2.5. Overlapping Functions and Interactions	17
2.6. An Evolutionary Link?	18
2.7. Acknowledgements	22
2.8. Materials and Methods	23
2.9. References	23
3. Assessing the Redundancy of Ethylene Receptor Function	29
3.1.0. Introduction	29
3.1.1. The <i>Arabidopsis</i> Receptor Family	30
3.1.2. Mode of Action and Implications of Ethylene Binding	33
3.1.3. Bridging the Gap: From Ethylene Perception to Ethylene Effect	35

3.2. Methods	36
3.2.1. Plant Material	36
3.2.2. Microscopy and Photomicrography	36
3.2.3. Sequence Analysis	37
3.3. Results and Discussion	38
3.3.1. The Dominant Ethylene Receptor Mutants Have Subtle, but Distinct Phenotypes	38
3.3.2. The Sequence Relationships and their Biological Relevance	48
 4. The Evolutionary Character of Plant Subtilisin-Like Proteases	 53
4.1. Introduction	53
4.2. Methods	60
4.2.1. Sequence Data	60
4.2.2. Phylogenetic Reconstructions	60
4.2.3. Structure Predictions	62
4.3. Results and Discussion	63
4.3.1. The Plant Subtilisin Proteases	63
4.3.2. Assessing the Properties of Sequence Groups	91
4.3.3. The Origins of Plant Pyrolysins	104
4.3.4. Towards a Global Model of Sequence Evolution	110
4.4. References	114
 5. General Conclusions	 113
 6. References	 115
 7. Appendices	 123

List of Tables

Table 1. Parameters Used for Phylogenetic Analysis	62
--	----

List of Figures

Figure 1: Aerenchyma and xylem are obvious features in the root of maize.	7
Figure 2: The developmental course of tubular structures proceeds along a similar developmental course in various plant structures	9
Figure 3: Domain organization of the plant ethylene receptor protein.	32
Figure 4: 4-week old ethylene receptor mutants.	40
Figure 5: Ethylene receptor mutants at 12 weeks of age.	42
Figure 6: Xylem anatomy of the dominant ethylene receptor mutants.	47
Figure 7: Protein sequence relationships of the plant ethylene receptors.	50
Figure 8: The structure of subtilisin-like proteases.	56
Figure 9: Sequence relationships of the plant subtilisin-like sequences.	65
Figure 10: Unrooted radial tree of plant-pyrolysin sequence relationships based on Neighbor-Joining analysis.	67
Figure 11: Radial unrooted tree of plant pyrolysin sequences based on a Bayesian inference of sequence relationships.	69
Figure 12A: The Arabidopsis pyrolysin family. Inference of sequence relationships based on Neighbor-Joining	78

analysis.

Figure 12B: The Arabidopsis pyrolysin family. Inference of sequence relationships based on Maximum-Likelihood analysis.	80
Figure 13A: The rice pyrolysin family. Sequence relationships based on NJ analysis.	82
Figure 13B: The rice family of pyrolysin subtilisin-like proteases as inferred by ML.	84
Figure 14: The tomato pyrolysin family as inferred by NJ analysis.	86
Figure 15: Hypothetical ancestral state of the plant subtilisin-like protease family prior to the monocot-dicot divergence.	90
Figure 16: Structural modelling of the plant subtilisin-like proteases.	94
Figure 17: The sequence relationships of the Arabidopsis subtilisin-like proteases can be simultaneously traced by physicochemical features.	98-102
Figure 18: Maximum-Likelihood estimate of sequence relationships of various members of the pyrolysin sub-family of subtilisin-like serine proteases.	107
Figure 19: The evolution of multi-protein families can be modelled through a spherical geometry.	112

List of Abbreviations

aa	amino acid
AIR	Auxin Induced Root
ALE	Altered Leaf Epidermis
ARA	<i>Arabidopsis</i>
B	Bayesian
BLAST	Basic Local Alignment Search Tool
Ca ²⁺	calcium ion
D	Aspartic acid
EIN	Ethylene Insensitive
ERS	Ethylene Response Sensor
ETR	Ethylene Response
EMBOSS	European Molecular Biology Open Software Suite
GAF	cGMP-specific and cGMP-stimulated phosphodiesterases,
<i>Anabaena</i>	Andenylate cyclases and <i>Escherichia coli</i> FhIA
H	Histidine
LIM	<i>Lilium</i> Meiotic
MCMC	Markov Chain Monte Carlo
MITE	Miniature Inverted Transposable Element
ML	Maximum Likelihood
N	Asparagine
NJ	Neighbor-Joining
NMR	Nuclear Magnetic Resonance
O ₂	molecular oxygen
ORF	Open Reading Frame
PAM	Point Accepted Mutation
PC	Protein Convertase
PCD	Programmed Cell Death
S	Serine
SDD	Stomatal Density and Distribution
SBT	Subtilisin-like serine protease (tomato)
SEM	Scanning Electron Microscope
SKI	Subtilisin/Kexin Isozyme
SLN	Synthetic Lethal of N-end rule
SREMBP	Sterol Regulatory Element Binding Protein
TE	Tracheary Element
TED	Tracheary Element Differentiation
TMP	Tomato Meiotic Protease
TUNEL	Terminal Deoxynucleotidyl transferase-mediated dUTP Nick-End Labelling
WAG	Whelan and Goldman
XSP	Xylem Serine Protease

List of Appendices

Appendix 1: Ethylene receptor protein sequences	123
Appendix 2: Protein sequence alignment of plant ethylene receptors	125
Appendix 3: Subtilisin-like protein sequences used	129
Appendix 4: Protein sequence alignment of plant pyrolysin proteases	133
Appendix 5: Protein modelling templates	140
Appendix 6: Protein sequence alignment of diverse pyrolysin subtilases	142

1. General Introduction

The use of *Arabidopsis thaliana* L. as a model genetic system in plants has substantially increased our understanding of basic plant biology. The utility of this organism in studying diverse features of plants continues to grow. It has recently been proposed that this organism can serve as an excellent model system in which to study the developmental physiology and genetic regulation of wood (secondary xylem) formation (Chaffey *et al.*, 2002). Xylem, the tissue through which water is transported in vascular plants requires unique modifications to the individual cells comprising this tissue. These modifications include the generation of the distinctive hydrophobic cell wall, and the autolytic programmed cell death of these individual vascular elements. The functional structure is a column of cell corpses (tubes) through which water ascends. Section 2. describes the current understanding of the specific developmental events in xylem development. It also presents evidence that a similar developmental theme is observed in other plant structures. It suggests a hypothesis that this theme is a more universal feature of plant form. This section was published in the April 2003 issue of the journal *Trends in Plant Science* (volume 8 p.159-164).

A microarray experiment was performed that identified a series of genes potentially differentially regulated during the generation of secondary xylem fibers in *Arabidopsis* (S.Regan, unpublished data). Several interesting candidates were identified and, based on their putative identities and potentially interesting functions in the regulation of xylem development, selected for further analysis.

Two of these candidates form the basis of the experimental work done here. The first of these is a receptor for the plant hormone ethylene. Ethylene is a small gaseous hydrocarbon and is one of the oldest known plant hormones. It has been implicated in the regulation of a number of diverse features in plant development and environmental response (Abeles *et al.*, 1992). The involvement of ethylene in xylem development is poorly characterized but a number of experiments indicate that it does play a role in xylem differentiation (Roberts and Miller, 1981). The ethylene signal transduction pathway has been characterized at the genetic level and 5 loci are believed to play a role in ethylene perception. The number of receptors is interesting because it would be expected that the different receptors are each responsible for the different ethylene effects we observe in plants. The picture that has emerged from the characterization is a complex one. Dominant mutants do not show great phenotypic differences between the different forms while recessive mutants only produce a phenotype when multiple receptor mutants are observed (Hua and Meyerowitz, 1998). This has led some authors to propose that ethylene receptor functions are redundant (Hua and Meyerowitz, 1998; Tieman *et al.*, 2000). Our analysis here investigates the premise that receptors are functionally redundant by using both physiological and molecular sequence analysis.

The redundancy of gene and protein function, particularly recently following gene duplication events is an area of active research in molecular evolution (Gu *et al.*, 2002). The last section of this work is an attempt to characterize a large family of proteases in plants at the level of molecular

sequence evolution. One specific protease, a member of the subtilisin-like proteases was identified in the microarray experiment. Proteases might be involved in xylem development in a number of interesting ways, such as molecular triggers or downstream effectors of Programmed Cell Death (PCD), or possible by acting as processing enzymes of peptide hormones. The characterization of 102 plant subtilisin-like protease sequences was performed to assess the sequence diversity and attempt to relate this diversity to functional roles. While the prediction of functional roles may currently be beyond informatic techniques, it is hoped that improved characterizations of both proteins and their evolutionary history will enable us to explore the evolutionary potential of peptide sequences. As an initial strategy, I suggest that raw sequence relationships might prove useful as a way of classifying particular, evolutionarily tractable, features of proteases, enzymes, or proteins in general.

2. How Plants Make Tubes

2.1. Abstract

The plant body requires the transport of various materials over large distances. Two cell types that bear a striking resemblance morphologically are the cells specialized for water transport and those responsible for the transport of oxygen: xylem and lysigenous aerenchyma, respectively. Each of these cell types

undergoes programmed cell death and cellular autolysis, resulting in the production of a functional space within the plant body. The major morphological difference observed is the presence of the lignified secondary wall in water-conducting tissues. The prevalence of tubular structures in other plant tissues suggests that the ability to form spaces through cellular autolysis is a fundamental paradigm in plant development and evolution.

2.2. Introduction

The evolution of large, multicellular organisms was accompanied by a broad expansion of cellular diversity. As organisms increased in size, the need for long-distance transport of materials increased substantially. In many organisms, this takes the form of hollow, tubular structures that are formed through various processes in multicellular development. In plants, long-distance transport, particularly through the vascular tissues, has been studied for many years. We observe that at least two cell types participating in long-distance transport, during the course of normal differentiation, produce similar cellular morphologies. The functional cell corpses of water-conducting cells and certain intercellular air spaces are morphologically similar. This similarity is characterized chiefly by the execution of a programmed cell death (PCD) program that is coordinated between adjacent cells within the plant body and has been previously discussed [1,2]. Higher plants have for a long time been known to make use of cell suicide mechanisms at various times and in various tissues [3]. Examples of this include

the hypersensitive response, in which a localized group of cells undergo PCD in response to an invading pathogen, thereby limiting the spread of infection, as well as in the senescence of numerous plant organs. During the development of the primary water-conducting cells found in plants, collectively called tracheary elements (TE's), the TE precursor cell initiates a cell-death program that results in the mature water-conducting cell. Only recently however, have the molecular features of PCD in TE differentiation received significant experimental attention.

The animal models of PCD have served as adequate templates upon which generalized features of programmed cell death have been identified. These features are important in that they provide a means of distinguishing programmed versus accidental cell death. Characteristics such as nuclear DNA fragmentation, and some form of cell corpse processing have been shown to occur in the majority of types of PCD investigated to date [4]. Incidental to these general features, it can be expected that plants have evolved specific features, characteristic to the unique developmental programs that have been observed in the various instances of plant PCD. Two instances of plant PCD that resemble each other at a gross cellular level are TE formation and lysigenous aerenchyma (air space) formation (Fig. 1). The development of both structures is roughly parallel in that they proceed by the following path; an undifferentiated parenchymatous cell undergoes cell expansion, and a cell death program that culminates in the autolysis of the cell leaving a cell corpse that functions in the transport of either a liquid (TE) or a gas (aerenchyma)(Fig. 2). The major distinction between these structures is the presence of a highly developed

Figure 1: Aerenchyma and xylem are obvious features in the root of maize. Cryo-planed SEM view of the transverse face of a corn root. (a) Overview showing numerous air spaces in the root cortex of corn. The xylem vessels are arranged in a ring in the central steele. Two of the vessels (at 12 o'clock and 10:30) are embolized (air-filled) while the remaining six are water-filled. (b) Close-up of the large aerenchyma to the left of the steele in (a). Cortical aerenchyma are often irregularly shaped, air-filled spaces that allow for the transport of oxygen in roots. (c) Two vessels shown from (a). The embolized vessel (left) was filled with air when frozen, while the water-filled one (right) was not.

Figure 2: The developmental course of tubular structures proceeds along a similar course in various plant structures. This course has been divided into 5 stages for comparison: Progenitor tissue (i), cell expansion (ii), the initiation of developmental fate and cell death (iii), programmed cell death (iv), and functional tissue (v). (a) The development of a tracheary element is marked most obviously by the development of the hydrophobic, degradation-resistant, secondary cell wall (yellow). (b) Large intercellular aerenchyma are often derived from the co-ordinated autolysis of adjacent cells that will allow for O₂ transport in O₂-poor environments. (c) The development of the female reproductive structure of mosses and ferns, the archegonium, involves the co-ordinated dissolution of the cell file that lies above the female gamete (white). At maturity this hollow space is the passage through which the sperm will access the egg.

secondary cell wall in tracheary elements and the absence of any obvious remnant cell wall structure in aerenchyma. The common cell morphology of these structures could provide clues to any characteristics of PCD that might be shared between these cell types. We propose that a detailed comparison at the cellular and molecular levels could test hypotheses regarding particular evolutionary relationships between these cell types.

2.3. Tracheary element formation

The accepted theory for the ascent of water in higher plants requires a highly developed vascular system in which water is conducted from the basal portion of the plant to the aerial portions. This vasculature supports a continuous water column within the plant body, marking the path of water flow. Plants have evolved a variety of specialized cell types within which the water column is maintained. Although these cells vary with respect to particulars of form, from xylem vessels to tracheids, both the general functions and developmental processes are shared. For this reason, all water-conducting cells will be referred to as tracheary elements (TE's). This is not to trivialize any functional disparities that exist between these cells *in vivo*, but rather to emphasize the common developmental program that is thought to be shared. Another simplification that will be used is that the development of TE's will be discussed irrespective of their origin within the plant body. That is, TE's of both apical and cambial meristematic origin will generally be considered similarly. This is probably an

oversimplification, particularly in specific developmental aspects such as secondary cell wall deposition. However, a detailed characteristic description of primary and secondary TE's is beyond the scope of this review.

The differentiation of TE's has been extensively studied *in vitro* in the *Zinnia* mesophyll system, developed >20 years ago [5]. Isolated TE's are complete with secondary cell walls and have been shown to undergo a PCD program [6]. The development of TE's in this system has been divided into three distinct stages. Stage 1 corresponds to the dedifferentiation phase and results in a pluripotent cell. Stages 2 and 3 of *Zinnia* TE development mark the formation of a mature TE from that of a pluripotent precursor. These stages best represent the situation of a developing TE *in vivo*, where no dedifferentiation occurs. Stage 2 marks the early redifferentiation events during TE formation. A great deal of recent work has attempted to characterize the genetic controls active during TE differentiation [7,8]. Although many genes have been identified as being differentially expressed, few have been extensively investigated in terms of specific function. For example, an activity has only recently been identified for the TED4 gene product. TED4 protein has been shown to inhibit endopeptidase activity in the extra-cellular matrix [9]. This might be an important finding in light of the central role of proteolysis in most forms of PCD.

Proteolytic activity has been shown to be a centerpiece of PCD in most systems studied. Early work in animal apoptosis identified a series of cysteine-dependent aspartic-specific proteases (caspases) that were initially thought to be the crucial initiators and effectors of apoptosis [4]. In recent years, the role of

caspases in triggering cell death has been minimized [10], but they remain important players in animal cell-death execution. No plant caspases have been identified to date [11], but the activity of caspase-like proteins has been demonstrated in cell death studies of tomato and tobacco cell suspension cultures. Two groups [12,13] have recently induced plant cell death with chemical insults known to induce apoptosis in animal cells. Cell death was subsequently blocked by the addition of caspase-specific inhibitors. These findings imply that although plants might not possess caspases, they do contain proteases that might perform similar roles *in vivo*. Coupled to this is the finding that both serine and cysteine proteases might be involved in the differentiation of *Zinnia* TE's *in vitro* [14]. In terms of cell death initiation, recent attention has shifted towards the mitochondrion and cytochrome c release as a potential trigger of PCD [10,15]. Although the current evidence implicates this organelle as being involved in TE PCD, its role in a central triggering mechanism is inconclusive [15].

The accumulation of hydrolytic enzymes, including proteases, DNAses and RNAses occurs within the vacuole of *in vitro Zinnia* TE's late in stage 2 of development and into stage 3 [16]. The beginning of stage 3 marks the developmental checkpoint where the cell commits to terminal differentiation. The major morphological marker of this stage is the presence of the lignified secondary cell wall.

Some time after completion of the secondary wall, the large central vacuole ruptures, releasing the store of hydrolytic enzymes into the cytoplasm [6]. The

events immediately following tonoplast rupture have been studied by Hiroo Fukuda and colleagues [17]. Sequential images of single developing TE's just before and immediately following vacuole collapse were observed by confocal-scanning laser video microscopy. Slightly before tonoplast rupture, the nucleus appears flattened at the cell periphery. This is presumed to be caused by the pressure exerted on it by the intact vacuole. Just after vacuole collapse, the nucleus assumes a more spherical shape that indicates a release of vacuolar pressure. A gradual decrease in nuclear staining is noticed after vacuole rupture. DNA degradation is a hallmark of PCD and has been demonstrated both *in vitro* [6] and *in vivo* [18] in developing TE's. Nucleic acid degradation here was presumed to occur via the release of endonucleases previously contained within the vacuole [17]. The nuclear membrane remained intact after DNA degradation and the nucleus itself was not fragmented.

Andrew Groover and Alan Jones [19] have proposed a model in which secondary cell wall synthesis is coordinated with triggering vacuolar collapse and subsequent cell death. Two key pieces of evidence were used in constructing this model. They identified a Ca^{2+} influx as a cell-death-triggering event. They also detected the secretion of a 40-kDa protease late in stage 3 of differentiation. Based on these and other results, Groover and Jones have proposed the following model. As secondary wall synthesis nears completion, a protease is released into the extracellular matrix, where it initiates a signal to release Ca^{2+} into the cell. The calcium influx then triggers vacuole rupture and cell death. This model adequately explains some of the features of PCD in this system, but

two key steps remain unresolved. First, what is the nature of the signal that triggers Ca^{2+} influx and what is its relationship to the secreted protease? Second, what is the nature of the mechanism by which Ca^{2+} triggers cell death? Does it directly affect vacuole rupture or are there intermediate steps that occur? These questions will require further study to elucidate the mechanism of cell death in developing TE's.

The *Zinnia* system has revealed a great deal about the differentiation of TE's but there has also been work done in *in vivo* systems. Proteolytic activity in relation to xylogenesis has been studied both in trees [20], and in *Arabidopsis*. Eric Beers' group [21] generated xylem-specific cDNA libraries from *Arabidopsis* hypocotyls and subsequently isolated two cysteine proteases and one serine protease that were found to be specifically expressed in xylem. It remains to be determined whether the enzymes coded for by these transcripts play important roles in TE formation. Studies such as these will play an important part both in confirming the findings from the *in vitro* systems as well as in determining effects on TE formation which are only present *in vivo*, such as cell positioning and cell-cell contact.

2.4. Lysigenous aerenchyma share developmental characteristics with TE's

The development of air spaces within the plant body is believed to be an adaptation to deal with conditions of low oxygen (hypoxia) [22]. Many plant species can be induced to form these structures, particularly when growing under

flooded soil conditions. Aerenchyma can form in two general ways, through schizogeny (cell separation) or lysigeny (cell death through lysis) [1].

Schizogenous aerenchyma are formed through a very different mechanism than lysigenous aerenchyma and their development is not considered here further.

Lysigenous aerenchyma form through a mechanism similar to that observed in TE's, namely PCD and cellular autolysis. Development of lysigenous aerenchyma with respect to PCD has received much less attention than PCD studied in tracheary elements. Early studies of aerenchyma formation focused on induction or on structural and ultrastructural aspects. Exogenous ethylene was found to mimic the effects of flooding, including air-space formation, in the root cortex of sunflower [23]. Using electron microscopy, the loss of tonoplast integrity was identified as an early event of aerenchyma development [24]. More recent work has concentrated on cellular mechanisms of aerenchyma formation. Malcolm Drew and colleagues [25] studied ethylene signal transduction and the role of intracellular calcium on aerenchyma formation in maize. They tested the effects of a variety of agents known to have an effect on cytosolic Ca^{2+} such as caffeine and thapsigargin. Both are known to increase cellular Ca^{2+} . Treatment with these compounds increased cell death under both normoxic (20.6% O_2) and hypoxic (4.0% O_2) conditions. Treatment with the Ca^{2+} chelator EGTA had the opposite effect, showing almost no cell death. Hirofumi Uchimiya's group investigated the patterning of cell death in the root cortical cells of rice [26]. They found that cell death occurred in radial files in the cortex. They also showed that cell death was initiated in the middle of the cortex and cell death spread inward

and outward along the radial cell file, suggesting that cell to cell communication via a diffusible message was an important feature of cell death in this system.

The hallmark features of PCD have only recently been investigated in a lysigenous aerenchyma-forming system in corn roots. Several key features, including terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling (TUNEL)-positive nuclei, DNA laddering, cytoplasmic condensation, and the presence of seemingly intact organelles within membranous compartments were detected [27]. These compartments resemble the apoptotic bodies found in many types of animal apoptosis. This is a key study showing that specific features of a PCD program are present in the development of these structures.

One piece of evidence that might shed some light on the relationship between TE and aerenchyma development is the phenotype of the *gapped xylem* (*gpx*) mutant of *Arabidopsis* [28]. These mutants show abnormalities in xylem vessel development such that some vessels fail to properly lay down a secondary cell wall layer that is resistant to the hydrolytic enzymes released upon vacuole rupture. The resulting structure is strikingly similar to the normal appearance of lysigenous aerenchyma. Although this evidence points to similar mechanisms involved in the developmental programs of these two cell types it remains strictly at the level of cellular morphology, namely tubes. A more detailed comparison will await a finer characterization of PCD in lysigenous aerenchyma. Such a characterization should include both additional cytological, biochemical and molecular markers of PCD, such as vacuole rupture, and the induction of a specific proteolytic program. Could the TED genes play a role in

aerenchyma formation? In more general terms, to what extent do models such as Andrew Groover and Alan Jones' extend to lysigenous aerenchyma formation? Answers to these questions should help to construct more complete descriptions of the development of long-distance transport systems in plants.

2.5. Overlapping function and interactions

Numerous reports exist describing interactions between water and air-conducting spaces in plant bodies. The use of cryo-fixation and microscopic methods have shown that aerenchymatous spaces in corn roots are sometimes filled with water [29]. How widespread this occurrence has yet to be determined but the prospect of large spaces within plant organs acting as water reservoirs is an exciting possibility. The ability of *Spartina alterniflora*, a salt marsh grass, to regulate the water table by root water uptake [30] was presumed by the authors to be because of water storage in the root parenchyma cells. In light of the findings that large spaces within plant organs might act as water reservoirs [29], it alternatively might involve the storage of water in the large lacunar spaces present in *S. alterniflora*. This would clearly require further study but could have important implications for plant-water relations.

The inverse phenomenon has stirred controversy with respect to plant-water relations. Traditional views of water transport support the notion that embolised xylem vessels are permanently incapacitated, the ability to transport water is irreversibly shut off in the air-filled vascular strand. The possibility of embolism repair, which has been raised several times in recent years [31,32],

would require a mechanism to fill embolised vessels with water from the surrounding tissues. A mechanism for this has been proposed [31] in which water is forced into empty vessels by the positive pressure generated by the surrounding tissues. However, the idea of embolism repair remains unresolved [33,34].

Whether or not traditional views of transpiration theory are upheld, there remains a fundamental relationship between the water-conducting tissues and air spaces within the plant body. The schizogenously generated air spaces between the spongy mesophyll cells in the leaf serve as the site of evaporation generating the pull on the transpiration stream. Are there examples of lysigenously derived air spaces acting as the main site of evaporation driving transpiration? In plants in which the main photosynthetic organ is the stem, stomata are commonly found on this organ [35]. The stomatal pores frequently open up into intercellular spaces that resemble aerenchyma when seen in transverse section. What sort of developmental program generates these spaces? How do they physically interact with the conducting xylem tissue? Answers to these questions might help elucidate the functional relationships between air-filled and water-filled spaces within the plant body.

2.6. An evolutionary link?

The close resemblance of lysigenous aerenchyma to the water-conducting cells in plants makes it tempting to speculate that these are homologous structures.

This homology is akin to the serial homology Richard Owen used in describing the relationship between any two vertebrae within a single animal [36]. The main difference between these two examples being the level of biological organization under consideration. Two scenarios present themselves when considering a relationship of homology. First, that water-conducting cells are ancestrally derived from cells originally responsible for the transport or storage of gasses. A highly developed, lignified secondary wall arose as an improved system of liquid conduction. Air spaces are a common feature of aquatic plants [37]. It seems possible then that the aquatic ancestor of land plants possessed such a trait. Little direct evidence currently exists as to whether this was the case. The fossil record is strikingly poor during the period in which plants colonized land [38]. The earliest land plants on record, such as *Cooksonia*, show presumed water-conducting cells with spiral thickenings reminiscent of modern tracheids [35]. Although air spaces are common in some groups of extant algae [39], they have not to our knowledge been reported in the lineages believed to be the closest relatives to the land plants; the Coleochaete's and Charales [40]. The inverse scenario of homology, where aerenchyma are derived from previously existing water-conducting cells is unlikely, because most cases of aerenchymatous tissue are derived from cortical cells, generally removed from the vascular tissues. Aerenchyma development also seems to have been a strategy that has arisen independently many times, as evidenced by the diversity of tissues within which they have been identified. The most common of these is cortical tissue but they

have also been reported in some secondary tissues including phellem [41] as well as occasionally being derived within vascular tissues [37].

The alternative hypothesis to a homologous relationship between these structures is that each of these has arisen independently. Homoplasy is the term used to describe two characters whose similarity derives from convergence as opposed to common descent. Homoplasy in relation to plant morphology has been discussed [42]. We believe that the ability to conduct fluids (gases or liquids) in plants is easily generated by the development of tubes derived from cellular autolysis. In addition to the evidence mentioned supporting the independent derivation of aerenchyma in land plant lineages, there is evidence that water-conducting cells among land plants might not be homologous. The development of moss hydroids is significantly different from the vessels and tracheids of vascular plants [43]. Even among the different bryophyte lineages there appears to be fundamentally different mechanisms for generating the connections between successive water-conducting cells suggesting independent origins. Several other tubular structures derived from cellular autolysis exist in some plant groups, such as within the archegonial complexes of mosses and ferns [44]. In these species, the oogonium develops at the base of the female reproductive organ - the archegonium. The cells that lie above the oogonium, the neck canal cells, seem to undergo a coordinated autolytic program resulting in the hollow neck of the mature archegonium (Fig. 2c). This chamber fills with water, allowing the flagellated sperm of these species to swim to the female gamete. It seems unlikely (although not impossible) that this structure is directly

homologous to either of the other tube structures discussed. A more probable relationship to cells in higher plants is homology to reproductive structures in seed plants. PCD is reported to be crucial to proper development of the transmitting tissue in seed plants [45]. In particular, the degeneration of the synergids, the cells that allow physical entry of the pollen tube into the ovule has been correlated to a sharp increase in intercellular Ca^{2+} [46]. Once again we are confronted with a lack of sufficient molecular and cytological data with which to adequately test predicted relationships. In spite of this, the widespread occurrence of these structures in various plant architectures might indicate a more subtle relationship between them.

Recent theoretical discussions of parallelisms have acknowledged the possibility that certain inherited factors influence the generation of similar forms independently. The idea of 'canalized evolutionary potential' or 'underlying synapomorphy' is the concept that the potential to develop a characteristic arises early in an organismal lineage but that the character arises independently in the subsequent evolutionary lines [47]. In the present discussion, it seems likely that one of the essential characters necessary for the generation of hollow structures within a three-dimensional multicellular body is the rigid cell wall. Hollow spaces formed by lysigeny or schizogeny have been widely recognized in both plants and fungi for many years [39]. In animals, tubes seem to be formed through vastly different mechanisms, primarily by the migration of cells or cell layers. These often fold back on themselves to generate hollow structures through which materials are transported [48]. The cell walls possessed by the aforementioned

groups seem to be a necessary feature for withstanding higher pressures generated by the surrounding tissues, particularly when the hollow spaces are air-filled. Most of the hollow structures reported in the fungi seem to arise through cell separation and differential growth rather than cell lysis [39,49]. Which features of plant cells might favor the generation of lysigenous spaces? An important feature might be the degree to which plant cells can regulate their cytoplasmic connections through plasmodesmata. Through the creation of secondary plasmodesmata, or alternatively the sealing of connections by callose occlusions, plants might be particularly adept at regulating the spread of developmental cues through a particular cell file.

Evolutionary relationships are difficult to test. The comparative method will need to be applied at a number of different biological levels to discern any relationship. The subtle cellular features and their molecular controls are available to modern researchers and can now be used to assess potential evolutionary relationships. In spite of the difficulties associated with these kinds of predictions, we believe that the potential to form tubes through cellular lysis was an important characteristic necessary for the evolution and development of the modern plant form.

2.7. Acknowledgments

We thank Barry Flinn and Nigel Chaffey for critical reading of the manuscript, Lewis Ling and Brendan O'Malley for technical assistance with the cryo-SEM,

Dan McPhee for assistance in preparing the figures and the Natural Sciences and Engineering Research Council of Canada for financial support.

2.8. Materials and Methods

Corn plants *Zea mays* L. cv. Seneca Chief were grown in a greenhouse to ten weeks of age. Large nodal roots (i.e. the oldest) approximately 30 cm in length were removed with a straight razor and wrapped in moist paper towel for transport to the microscopy centre. A small, 5 mm piece was excised from the root approximately 20 cm from the root tip and frozen in liquid nitrogen. This piece was mounted on a stub with Tissue-Tek (Miles Inc., Elkhart, IN, USA) and placed back in liquid nitrogen. Frozen samples were planed with a cryo-microtome (CR200, Research and Manufacturing, Inc., Tuscon, AZ, USA) at -80°C. The sample was then transferred to the cryochamber under liquid nitrogen. The sample was etched by slight warming for a few seconds on the stage of the cryo-SEM (JEOL 6400, JEOL Ltd., Tokyo, Japan). The sample was lightly coated with aluminium and observed in the microscope at -170 °C at 7-10 kV.

2.9. References

- 1 Drew, M.C. *et al.* (2000) Programmed cell death and aerenchyma formation in roots. *Trends Plant Sci.* 5, 123-127

- 2 Jones, A. (2001) Programmed cell death in development and defense. *Plant Physiol.* 125, 94-97
- 3 Grey, J. and Johal, G.S. (1998) Programmed cell death in plants. In *Arabidopsis Annual Plant Reviews* 1, 360-394 (Anderson, M. and Roberts, J.A., eds)
- 4 Danon, A. *et al.* (2000) Plant programmed cell death: a common way to die. *Plant Physiol. Biochem.* 38, 647-655
- 5 Fukuda, H. and Komamine, A. (1980) Establishment of an experimental system for the study of tracheary element differentiation from single cells isolated from the mesophyll of *Zinnia elegans*. *Plant Physiol.* 65, 57-60
- 6 Groover, A. *et al.* (1997) Programmed cell death of plant tracheary elements differentiating in vitro. *Protoplasma* 196, 197-211
- 7 Lorenz, W.W. and Dean, J.F.D. (2002) SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (*Pinus taeda*). *Tree Physiol.* 22, 301-310
- 8 Demura, T. *et al.* (2002) Visualization by comprehensive microarray analysis of gene expression programs during transdifferentiation of mesophyll cells into xylem cells. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15794-15799
- 9 Endo, S. *et al.* (2001) Inhibition of proteasome activity by the TED4 protein in extra-cellular space: a novel mechanism for protection of living cells from injury caused by dying cells. *Plant Cell Physiol.* 42, 9-19
- 10 Jones, A. (2000) Does the plant mitochondrion integrate cellular stress and regulate programmed cell death? *Trends Plant Sci.* 5, 225-230

- 11 Lam, E. and del Pozo, O. (2000) Caspase-like protease involvement in the control of plant cell death.
Plant Mol. Biol. 44, 417-28
- 12 De Jong, A.J. *et al.* (2000) Chemical-induced apoptotic cell death in tomato cells: involvement of caspase-like proteins. *Planta* 211, 656-662
- 13 Elbaz, M. (2002) Constitutive caspase-like machinery executes programmed cell death in plant cells. *Cell Death Diff.* 9, 726-733
- 14 Ye, Z. and Varner, J. (1996) Induction of cysteine and serine proteases during xylogenesis in *Zinnia elegans*. *Plant Mol. Biol.* 30, 1233-1246
- 15 Yu, X-H. *et al.* (2002) Mitochondrial involvement in tracheary element programmed cell death. *Cell Death Diff.* 9, 189-198
- 16 Fukuda, H. (1997) Tracheary element differentiation. *Plant Cell* 9, 1147-1156
- 17 Obara, K. *et al.* (2001) Direct evidence of active and rapid nuclear degradation triggered by vacuole rupture during programmed cell death in *Zinnia*. *Plant Physiol.* 125, 615-626
- 18 Mittler, R. and Lam, E. (1995) *In situ* detection of nDNA fragmentation during the differentiation of tracheary elements in higher plants. *Plant Physiol.* 108, 489-493
- 19 Groover, A. and Jones, A. (1999) Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. *Plant Physiol.* 119, 375-384

- 20 Iliev, I. and Savidge, R. (1999) Proteolytic activity in relation to seasonal cambial growth and xylogenesis in *Pinus banksiana*. *Phytochemistry* 50, 953-960
- 21 Zhao, C. *et al.* (2000) Exploiting secondary growth in *Arabidopsis*. Construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. *Plant Physiol.* 123, 1185-1196
- 22 Jackson, M.B. and Armstrong, W. (1999) Formation of aerenchyma and the processes of plant ventilation in relation to soil flooding and submergence. *Plant Biol.* 1, 274-287
- 23 Kawase, M. (1974) Role of ethylene in induction of flooding damage in sunflower. *Physiol. Plant.* 31, 29-38
- 24 Campbell, R. and Drew, M.C. (1983) Electron microscopy of gas space (aerenchyma) formation in adventitious roots of *Zea mays* L. subjected to oxygen shortage. *Planta* 157, 350-357
- 25 He, C. *et al.* (1996) Transduction of an ethylene signal is required for cell death and lysis in the root cortex of maize during aerenchyma formation induced by hypoxia. *Plant Physiol.* 112, 463-472
- 26 Kawai, M. *et al.* (1998) Cellular dissection of the degradation pattern of cortical cell death during aerenchyma formation of rice roots. *Planta* 204, 277-287
- 27 Gunawardena A.H. *et al.* (2001) Characterisation of programmed cell death during aerenchyma formation induced by ethylene or hypoxia in roots of maize (*Zea mays* L.). *Planta* 212, 205-214

- 28 Turner, S. and Hall, M. (2000) The *gapped xylem* mutant identifies a common regulatory step in secondary cell wall deposition. *Plant J.* 24, 477-488
- 29 van der Weele, C.M. *et al.* (1996) Water in aerenchyma spaces in roots. A fast diffusion path for solutes. *Plant Soil* 184, 131-141
- 30 Dacey, J.W.H. and Howes, B.L. (1984) Water uptake by roots controls water table movement and sediment oxidation in short *Spartina* marsh. *Science* 224, 487-489
- 31 Canny, M.J. (1998) Applications of the compensating pressure theory of water transport. *Am. J. Bot.* 85, 897-909
- 32 McCully, M.E. (1999) Root xylem embolisms and refilling. Relation to water potentials of soil, roots, and leaves and osmotic potentials of root xylem sap. *Plant Physiol.* 119, 1001-1008
- 33 Canny, M.J. *et al.* (2001) The cohesion theory debate continues. *Trends Plant Sci.* 6, 454-455
- 34 Cochard, H. *et al.* (2001) The cohesion theory debate continues. Response by Hervé Cochard, Thierry Améglio and Pierre Cruiziat. *Trends Plant Sci.* 6, 456-457
- 35 Edwards, D. (1993) Cells and tissues in the vegetative sporophytes of early land plants. *New Phytol.* 125, 225-247
- 36 Panchen, A.L. (1994) Richard Owen and the concept of homology. In *Homology, The Hierarchical Basis of comparative Biology*. (Hall, B.K., ed.), pp. 21-62, Academic Press

- 37 Arber, A. (1920) *Water Plants, a Study of Aquatic Angiosperms*, Cambridge University Press
- 38 Thomas, B. and Spicer, R. (1987) *The Evolution and Palaeobiology of Land Plants*, Dioscorides Press, Portland, Oregon, USA
- 39 Sifton, H.B. (1945) Air-space tissue in plants. *Bot. Review* 11, 108-143
- 40 Karol K.G. *et al.* (2001) The closest living relatives of land plants. *Science* 294, 2351-2353
- 41 Stevens, K.J. *et al.* (2002) The aerenchymatous phellem of *Lythrum salicaria* (L.): a pathway for gas transport and its role in flood tolerance. *Ann. Bot.* 89, 621-625
- 42 Bateman, R.M. (1996) Nonfloral homoplasy and evolutionary scenarios in living and fossil land plants. In *Homoplasy: The Recurrence of Similarity in Evolution* (Sanderson, M.J. and Hufford, L., eds), pp. 91-130, Academic Press
- 43 Ligrone, R. *et al.* (2000) Conducting tissues and phyletic relationships of bryophytes. *Philos. Trans. R. Soc. London, Ser. B* 355, 795-813
- 44 Doyle, W.T. (1970) *The Biology of Higher Cryptogams*, The Macmillan Company
- 45 Wu, H. and Cheung, A.Y. (2000) Programmed cell death in plant reproduction. *Plant Mol. Biol.* 44, 267-281
- 46 Tian, H. and Russell, S.D. (1997) Calcium distribution in fertilized and unfertilized ovules and embryo sacs of *Nicotiana tabacum* L. *Planta* 202, 93-105

- 47 Brooks, D. (1996) Explanations of homoplasy at different levels of biological organization. In *Homoplasy: The Recurrence of Similarity in Evolution* (Sanderson, M.J. and Hufford, L., eds), pp. 3-36, Academic Press
- 48 Hogan, B.L.M. and Kolodziej, P.A. (2002) Molecular mechanisms of tubulogenesis. *Nat. Rev. Genet.* 3, 513-523
- 49 Cairney, J.W.G. *et al.* (1988) Structural differentiation in maturing rhizomorphs of *Armillaria mellea* (Tricholomatales). *Nova Hed.* 46,1-25

3. Assessing the Redundancy of Ethylene Receptor Function

3.1.0. Introduction

The use of ethylene as a phytohormone by plants has been the subject of intensive study by physiologists for the past 100 years. Classical experiments identified ethylene as affecting fruit ripening, seed germination, senescence and abscission. Ethylene has also been implicated in various stress responses such as those observed during pathogen attack, drought and flooding (Kieber and Ecker, 1993). Ethylene is produced endogenously by plants and its biosynthetic

pathway has been well characterized. Despite the early identification of this gaseous compound as a plant growth regulator, attempts to determine its particular modes of action had only been moderately successful until the advent of modern molecular biological techniques.

3.1.1. The *Arabidopsis* Ethylene Receptor Family

Perception of an ethylene stimulus remained somewhat of an enigma until the generation of ethylene-insensitive mutants of *Arabidopsis* (Bleeker *et al.*, 1988) and the subsequent cloning of the first putative ethylene receptor gene *ETR1* (Chang *et al.*, 1993). Subsequent work (Hua, *et al.*, 1995; Hua *et al.*, 1998; Sakai *et al.*, 1998) identified a multi-gene family in *Arabidopsis* consisting of 5 members designated *ETR1*, *ETR2*, *ERS1*, *ERS2*, and *EIN4*. These five receptors have been shown to have significant sequence and structural similarity to the bacterial two-component histidine kinase receptors (Figure 3). In bacteria these proteins are responsible for sensing a wide range of environmental stimuli including host detection, nutrient detection and osmolarity sensing (Parkinson and Kofoed, 1992). Several eukaryotic analogues of these bacterial receptors have been identified including the SLN1 osmolarity sensor in *Saccharomyces cerevisiae* (Ota and Varshavsky, 1993). Both *ETR1* and *SLN1* have a similar structure, bearing a receiver domain at the carboxy-terminus (Chang and Meyerowitz, 1995). The other identified ethylene receptors in *Arabidopsis* do not all have this same structure. *ETR2* and *EIN4* share the same domain architecture as *ETR1*, possessing a C-terminal receiver domain. *ERS1* and

Figure 3: The ethylene receptors are members of the two-component receptor family. A) General organization of the two-component system. The two components are the sensor and the response regulator. Each of these proteins is comprised of two functional domains. The sensor has an amino-terminal input domain and a carboxy-terminal transmitter module. The response regulator has an amino-terminal receiver module and a carboxy-terminal output domain. An incoming stimulus is recognized by the input domain and a phosphorylation cascade is initiated from the transmitter module, through the response regulator, and ultimately to the stimulus response. Adapted from Parkinson and Kofoed (1992). B) The yeast osmolarity sensor SLN1, a eukaryotic “two-component” sensor. The receiver module (circle) is on the same peptide chain as the input (triangle) and transmitter (rounded rectangle) domains. C) The 5 *Arabidopsis* ethylene receptors are a family of eukaryotic “two-component” sensors. All members share the ethylene-binding (skinny ellipses), a GAF (cGMP-specific and cGMP-stimulated phosphodiesterases, *Anaebena* *Adenylate* cyclases and *Escherichia coli* *FhlA*; fat ellipses) domain, and a histidine kinase domain (rounded rectangle). Additional domains are present in some members but lacking in others. ETR1, EIN4, and ETR2 have a carboxy-terminal receiver domain (rectangle), similar to SLN1. EIN4, ERS2, and ETR2 have an additional amino-terminal domain (filled rectangle), suggested to be a targeting peptide (Bleecker, 1999). Locations of the dominant mutations in the ethylene-binding domains present in the mutant plants used in this work are indicated. B) and C) are adapted from Bleecker (1999) and Hua *et al.* (1998).

ERS2 both have the same N-terminal domains, the ethylene-binding, GAF, and histidine kinase domain but lack the receiver domain. The EIN4, ETR2, and ERS2 proteins also possess a short, amino-terminal sequence believed to be a targeting sequence (Bleeker, 1999). This sequence does not seem to be necessary for the localization of receptors in all cases. The ETR1 protein has been detected as bound to the membrane of the endoplasmic reticulum by biochemical fractionation and immunoelectron microscopy (Chen *et al.*, 2002) despite it's lack of an amino-terminal targeting sequence. The sub-cellular localization patterns of the other ethylene receptors are not known. The ethylene receptor genes are distributed among 3 different chromosomes in the *Arabidopsis* genome. *ERS2* and *ETR1* are located on chromosome 1, *ERS1* is located on chromosome 2 and *EIN4* and *ETR2* are found on chromosome 3. The receptor genes have some structural similarities. *ETR2*, *EIN4*, and *ERS2* all have a only a single intron located at the same position within the region that codes for the transmitter domain. *ETR1* and *ERS1* also share this intron, as well as two others within the same domain, and one within the GAF domain. *ETR1* also has a single intron within the receiver domain sequence that is not shared with other members that possess this domain (Hua *et al.*, 1998).

3.1.2. Mode of Action and Implications of Ethylene Binding

While initial results from mutagenic screens and sequence similarity to bacterial environmental sensors implied that ETR1 was an ethylene receptor, conclusive evidence that the ETR1 protein could bind ethylene did not come until the work of

Eric Schaller and Anthony Bleeker, (1995). They were able to detect bound ethylene when the ETR1 protein was expressed in transgenic yeast cells. No ethylene binding was detected in yeast lacking the transgene, as well as in yeast transformed with the mutant allele *etr1-1* that confers ethylene insensitivity in *Arabidopsis*. Purification of the ETR1 protein from *Arabidopsis* and from transgenic yeast also showed that ETR1 exists as a dimer when purified from membrane fractions (Schaller *et al.*, 1995). Three membrane-spanning, hydrophobic domains are predicted in the first 130 amino acid (aa) residues of the mature protein (Figure 3) and this is the region purported to be responsible for ethylene binding based on the location of mutations conferring ethylene insensitivity and the ethylene-binding assay performed in yeast. These mutations were generated by mutagenic screens for plants that failed to undergo the triple response (Bleeker *et al.*, 1988; Hua *et al.*, 1995; Roman *et al.*, 1995; Sakai *et al.*, 1998). The triple response is a classical ethylene response in plant seedlings where the root and hypocotyl are shortened and thickened and there is an exaggerated apical hook on the shoot in dark-grown plants exposed to exogenous ethylene. The *ers1-1* mutant was obtained in a different fashion. This mutant was generated by site-directed mutagenesis and subsequent transformation of the mutated *ERS1* gene into wild-type *Arabidopsis* plants via *Agrobacterium*-mediated transformation (Hua *et al.*, 1995). All of the mutations that occur within the ethylene-binding domain of these receptors confer dominant ethylene insensitivity to the mutant plants.

The remainder of the input domain contains a predicted GAF (cGMP-specific and cGMP-stimulated phosphodiesterases, *Anabaena* Adenylate cyclases and *Escherichia coli* EhIA) domain (Hall *et al.*, 2001). The biological function of the domain is not well understood but it is believed to be involved in cGMP binding and signaling. Early in ethylene perception studies it was also proposed that a transition metal would be used by a putative receptor at the ethylene binding site (Hall *et al.*, 2001). Work by Rodriguez *et al.* (1999) has demonstrated that this is in fact the case. A copper (I) ion is coordinated by ETR1 at the suspected ethylene binding site. This was verified by site-directed mutagenesis of the residues predicted to coordinate this ion and resulted in a complete loss of ethylene binding ability.

3.1.3. Bridging the Gap: From Ethylene Perception to Ethylene Effect

The previous century of study identified a host of plant responses and effects ascribed to ethylene. The previous decade has seen the dissection of the genetic basis underlying ethylene perception in plants. How the ethylene signal is interpreted by plant cells and in turn is manifested into the observed physiological responses we are familiar with remains the key problem in plant-ethylene physiology. Two observations based on the receptor mutants have complicated this problem. The dominant ethylene-insensitive mutants do not show dramatic phenotypes relative to one another (Hua *et al.*, 1995; Sakai *et al.*, 1998) and mutant phenotypes of the recessive loss-of-function mutants are apparent only in double, triple and quadruple receptor mutants (Hua and

Meyerowitz, 1998). These features have precipitated the hypothesis that the ethylene receptors fulfil largely redundant roles *in vivo* (Hua *et al.*, 1995; Hua and Meyerowitz, 1998; Tieman *et al.*, 2000). This work attempts to further investigate the phenotypic characteristics of the dominant receptor mutants in *Arabidopsis* as well as characterize the receptor sequence family.

3.2. Methods

3.2.1. Plant Material

Arabidopsis mutants *etr1-1*, *ers1-1*, *ein4-1*, and *etr2-1* were grown in a greenhouse under short-day conditions. Only one mutant line was planted per tray. These trays of mutant plants were distributed such that each mutant was exposed to the various range of environmental conditions present in this house. Observations of various morphological features were made regularly.

3.2.2. Microscopy and Photomicrography

The secondary xylem anatomy of 12-week old mutant plants was examined. Fresh cut hand-sections were mounted in a drop of saturated phloroglucinol (Fisher Scientific) in 18% hydrochloric acid on a microscope slide. These were observed by bright-field microscopy on an Olympus Vanox Universal Research Microscope. Photomicrographs were taken using Agfachrome RSX 50 colour slide film. Digital images were taken directly from the colour slides placed on a

light table with an 8-bit CCD camera (Alpha Innotech Corporation, San Leandro, CA).

3.2.3. Sequence Analysis

Protein sequences were extracted from Genbank (<http://www.ncbi.nlm.nih.gov>) based on sequence similarity to the *Arabidopsis* ethylene receptors. Four additional sequences were identified by BLAST (Altschul, *et al.*, 1990) searches querying various *Arabidopsis* ethylene receptor sequences against the whole-genome shotgun sequence of *Oryza sativa* ssp. indica (<http://www.ncbi.nlm.nih.gov/BLAST/Genome/PlantBlast.html>). One hit was discarded due to low similarity. The other two were inspected for open reading frames (ORF's) using ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>). Predicted ORF's were manually spliced and translated with transeq, part of the EMBOSS package (<http://bioinfo.pbi.nrc.ca:8090/EMBOSS/index.html>). The three remaining sequences were designated Rice 3, 4, and 5. Sequences used in the analysis are identified in Appendix 1. These were used in a phylogenetic sequence analysis performed in collaboration with Dr. Guy Drouin at the University of Ottawa, Ottawa, Canada. Full-length protein sequences were used to construct a Neighbor-Joining tree using the program MEGA v.2.1 (Kumar *et al.*, 2001). Rate heterogeneity was modeled by a gamma distribution. Deletion of gaps was pair-wise. 500 bootstrap replicates were used to assess tree reliability.

3.3. Results and Discussion

3.3.1. The dominant ethylene receptor mutants have subtle, but distinct phenotypes

The suggestion that the *Arabidopsis* ethylene receptors perform overlapping, or redundant functions *in vivo* (Hua and Meyerowitz, 1998) necessitated a more complete phenotypic characterization of these mutants. The dominant, ethylene-insensitive mutant plants were observed in detail at four and twelve weeks of age. Even young plants showed minor differences in phenotype (Figure 4). The *ein4-1* mutant shows a consistent leaf phenotype that is easily observed. The leaf margins of this mutant have a distinctive serrated character and this phenotype can be used to easily score this mutant among the different receptor mutants (S. Regan, personal communication). Aside from this phenotype, these younger plants appeared to be quite similar in general appearance. As the plants aged, they appeared to show quite different phenotypes. At 12 weeks of age, mutant plants had a much more extreme appearance (Figure 5). The *ein4-1* and *etr2-1* plants still appeared quite similar to each other. Both of these lines had developed very dark purple leaves. This characteristic is often seen in *Arabidopsis* plants exposed to environmental stresses (unpublished observations) and a role for anthocyanins (purple, plant flavonoid pigments) in general plant stress responses is well documented

Figure 4: 4-week old ethylene receptor mutants. These plants appear quite similar in phenotype. Note the serrated leaf margins of *ein4-1*.

Figure 5: Ethylene receptor mutants at 12 weeks of age. Notice the severe coloration of the *ein4-1* and *etr2-1*. *etr1-1* and *ers1-1* showed premature senescence of rosette leaves.

(Winkel-Shirley, 2002). What these roles might be is not well understood but a favored hypothesis is that they act as photoprotectants during leaf senescence (Feild *et al.*, 2001). One possibility as to the particular stress affecting these plants could be the light intensity or quality they were exposed to. These plants were grown in the winter months and light reflected by the snow-covered ground could possibly be adversely affecting them. It was also noticed that many of the *etr2-1* plants had serrated leaf margins. These were slightly less distinct than those observed in the *ein4-1* plants. The most distinctive feature between these older plants was the general difference in appearance of *etr1-1* and *ers-1* plants compared to the *ein4-1* and *etr2-1* plants. None of the *etr1-1* or *ers-1* plants showed the dark purple colouration present in the other two mutant lines. The *etr1-1* and *ers-1* plants were both green and showed no sign of increased red colouration. The *ers-1* plants had several other distinct phenotypic differences. The basal leaves of the rosette, i.e. the oldest leaves, showed a tendency to be prematurely senescing. Several leaves per plant appeared completely dead while others appeared to be in the process of senescing. This phenotype was not observed in *ein4-1*, *etr2-1*, or wild-type plants. Some *etr1-1* plants showed a mild version of this premature leaf senescence, while others appeared fully green. This phenotype is not completely surprising considering the important role ethylene is thought to play during senescence (Abeles *et al.*, 1992). The *ers-1* plants had one phenotype that was not observed in any of the other mutant lines. These plants all bolted (flowered) much earlier than the other mutants. At twelve weeks of age, all the *ers-1* plants had flowered and were producing

siliques. The *etr1-1* plants were just beginning to initiate flowers at this time, and some immature floral buds could be observed in the central portion of the rosette. Both the *ein4-1* and *etr2-1* plants did not visibly show signs of flowering for several more weeks. None of the phenotypes observed here can conclusively be ascribed to ethylene specifically. We may be observing secondary phenotypes resulting from general stress responses not directly brought about by ethylene insensitivity. Further study is necessary to determine if any of these phenotypes arise directly as a result of faulty ethylene perception or by secondary processes (pleiotropic effects). A hypothetical example may serve to highlight this important difference. Ethylene may directly participate in the signalling in the initiation of anomolous senescing of the basal rosette leaves. This early senescence may in and of itself trigger the early flowering seen in the *ers1-1* mutant. Here ethylene perception is only secondarily involved in the early flowering initiation. More detailed physiological characterization with respect to specific environmental stresses would help clarify which responses can be directly attributed to faulty ethylene perception. Despite the difficulties in ascribing specific functional roles to specific receptors in this analysis, one important feature is evident. The different phenotypes observed here suggest that the functional importance of ethylene perception is not equally shared among the different receptor isoforms observed here.

As an additional experiment to observe potential differences in the phenotypes of the receptor mutants, the internal hypocotyl xylem anatomy of the four receptor mutants was observed. The hypocotyl of *Arabidopsis* produces a

significant amount of secondary xylem (Chaffey *et al.*, 2002). This region can be divided into two structurally distinct regions when viewed in transverse section. The first phase of secondary xylem is produced in younger plants and consists of xylem vessels surrounded by parenchyma. As the plant ages, the second phase is initiated and is continuously produced. This is characterized by the continuous production of xylem vessels interspersed with mature xylem fibers. Initial observations suggested that the initiation of phase 2 xylem was temporally correlated with the onset of flowering. The xylem anatomy of the ethylene-receptor mutants was observed by staining the tissues with phloroglucinol, which selectively stains lignin, identifying regions in the tissue laying down hydrophobic secondary cell walls. Both xylem vessels and xylem fibers are readily identified by this method. No significant differences were observed in the xylem anatomy between the various receptor mutants (Figure 6). The general observation that the initiation of stage 2 xylem coincides with flowering was affirmed by this analysis. For the *etr1-1* and *ein4-1* mutants, there was little or no phase 2 xylem observed prior to the plant having flowered. Hypocotyls that were observed after flowering showed the presence of mature xylem fibers among the vessels. The *etr2-1* mutant showed the presence of a small amount of phase 2 xylem even prior to observed floral initiation. These results affirm the possibility that the transition to flowering acts as a trigger for the initiation of phase 2 xylem. Further experimentation will be necessary to extend the relationship of floral initiation to phase 2 xylem initiation beyond one of temporal correlation. One approach could be to utilize *Arabidopsis* ecotypes that require vernalization to initiate flowering.

Figure 6: Xylem anatomy of the dominant ethylene receptor mutants.
Transverse sections of the *Arabidopsis* hypocotyl stained with phloroglucinol.
Figures at left have not flowered (not bolted) and those at the right have
flowered (bolted).

Non-bolted

Bolted

ers-1

etr1-1

ein4-1

etr2-1

These ecotypes will continue to grow for long periods of time without flowering if they are not cold treated. Examination of the xylem anatomy of both vernalized and non-vernalized treatments could test the hypothesis that flowering triggers xylem fiber formation.

3.3.2. The Sequence Relationships and their Biological Relevance

The analysis of 60 protein sequences from various plant species, predicted to be ethylene receptors based on sequence similarity to the experimentally verified receptors, is presented (Figure 7). The most obvious feature of the phylogenetic tree is the presence of three groups. These three groups are identified as the ERS1, ETR1, and ETR2/EIN4/ERS2 groups. These names reflect the association of the *Arabidopsis* sequences of the same names. In a few instances the names of other sequences does not correspond to the group in which it is found. This is assumed to be due to a cursory examination relative to established receptor sequences done by the author of the sequence, prior to database submission. It is assumed that the sequence relationship shown here has more biological relevance than the name given by the author. The general topology of this tree agrees with previous work done with a much smaller dataset that examined the sequence relationships based only on the ethylene-binding domain (Bleeker, 1999). There is a relatively high degree of similarity between the ERS and ETR1 groups as evidenced by the short branch lengths both within each group and between the two. This contrasts with the large distances

Figure 7: Protein sequence relationships of the plant ethylene receptors. The receptors fall into 3 distinct groupings, the ERS1 group, the ETR1 group, and the ETR2/EIN4/ERS2 group. The exact significance of these groups is not known, but this clustering suggests that selective constraints are maintaining these as distinct.

observed between these two groups and the third group, ETR2/EIN4/ERS2. There is also a large amount of sequence divergence within this third group. Despite this, each major group is distinct and the relationships are generally well-supported. Within two of these groups (the ETR1 and ERS1 groups) it appears as if a relatively strong, species-level phylogeny exists. For example, at the bottom of the ERS group lie three monocot sequences *OryzaERS*, *BananaER*, and *PhalaETR1* (orchid). A relatively accurate dicotyledonous tree lies above these correctly grouping Solanaceous (Potato, Tomato, Tobacco), Brassicaceae (*Brassica*, *Arabidopsis*) and other plant families in meaningful ways. This is a strong indication that these proteins are orthologous, i.e. they are related by descent through speciation and they perform the same function in different species. If this is the case the next logical question pertains to the nature of that function. Assessing function based on sequence similarity is not generally acceptable unless independently gathered information can assess this. In the present work, observations of mutant phenotypes suggests that there are important differences in signaling through the ethylene receptors. The division of the mutant phenotypes into two broad classes, one that is characterized by enhanced anthocyanin synthesis and late flowering and the other by premature senescence and early flowering by itself implies different functional roles. An alternative may be that the observed differences may be due to differential insensitivity to ethylene between the mutant plants used here. At present, only ETR1 has been conclusively shown to bind ethylene (Schaller and Bleeker, 1995). The different mutants used here have lesions in different positions within

their proposed ethylene-binding domains (Figure 3). It is possible then that these different lesions interfere with ethylene-binding to different degrees. Binding assays with the different receptor mutants could detect differences in ethylene binding and therefore test whether differential insensitivity is possible in these mutants.

The phenotypic division corresponds to the major division in sequence relationships observed i.e. that ETR1 and ERS1 are distinct from ETR2/EIN4/ERS2. This is an interesting finding that bears further investigation. One approach could be to perform sequence analysis on selected domains within the receptors, in particular the GAF and histidine kinase domains, as they are common to all the members. These are believed to be responsible for the downstream signalling of ethylene detection. If the same major divisions were observed as those seen with the entire protein, this would suggest that differential signaling was taking place. The use of full-length protein sequence in this analysis may have introduced an artifactual distinction between the ETR1 and ERS1 sequence groups as the former possessed a receiver domain and the latter did not. Further work using only domains shared among all the receptors could address this. The grouping of ERS1 sequences with the ETR1 sequences and not with other members that lacked a receiver domain suggests however that this may not be an overriding factor. The species-level phylogeny that is apparent within both the ETR1 and ERS1 groups suggests that the sequences within these groups are orthologs. Based on this it seems likely that some

elements of functionality would be shared between the different species' receptors that fall within the same sequence group.

4. Evolutionary Character of Plant Subtilisin-Like Proteases

4.1. Introduction

The subtilisin-like serine proteases are a large group of enzymes that have garnered a considerable amount of experimental attention. They represent one of the most well-characterized groups of enzymes on several research fronts. There are two primary focal points of interest that have driven research of these proteases. The first is the extensive protein engineering work that has been performed using several of the bacterial members of this family, primarily driven by their use in industrial applications (Bryan, 2000). This work has led to a detailed characterization of both the physical and kinetic properties of these enzymes (Creighton, 1993). The second, and the one that chiefly concerns the work presented here, is the relatively recent establishment of these enzymes as key players in the maturation pathway of peptide signals in eukaryotes. Best characterized in mammals, small peptides have been shown to act as important inter- and intra-cellular messages in a wide variety of biological processes (Bergeron *et al.*, 2000). The study of peptide signals in plants has identified only a few definite candidates but it is generally believed that small peptides play a major role in plant inter-cellular communication (Takayama and Sakagami, 2002).

All serine proteases are believed to act through a similar catalytic mechanism. Three conserved active site residues broadly define these proteases and are known generally as the catalytic triad. These three residues, an Aspartic acid (D), a Histidine (H), and the Serine (S) residue for which these enzymes are named, participate within the active site in the hydrolysis of peptide bonds. A fourth conserved residue, an Asparagine (N), resides in the region known as the oxy-anion hole. This region is responsible for the correct orientation of the target peptide bond within the active-site (Creighton, 1993). The linear arrangement of these residues in the primary sequence of these enzymes is an important feature of serine proteases. The subtilisin-like enzymes' primary sequence arrangement of D, H, and S (Figure 8) distinguish them from the other serine proteases, the trypsin family, whose order of active site residues is H, D, and S. This difference in linear order is accompanied by a significantly different tertiary structure of these two families and illustrates an example of apparent evolutionary convergence at the molecular level (Creighton, 1993; Siezen and Leunissen, 1997).

The subtilisin superfamily itself has been characterized, and 6 families have been suggested by the scheme of Siezen and Leunissen (1997). These are: 1) the subtilisin family; 2) the thermitases; 3) the proteinase K family; 4) the lantibiotic peptidases; 5) the kexin family; and 6) the pyrolysins. The first four of these family's members are primarily of prokaryotic origin. The fifth family, the kexin group, contains both fungal and metazoan enzymes that have been the subject of considerable interest. A group of metazoan enzymes in this class, the

Figure 8: The structure of subtilisin-like proteases. The linear order of reactive-site residues is a defining feature of the subtilisin-class of serine proteases. A) Schematic organization of a bacterial subtilisin. The arrangement of the active site residues, aspartic acid (D), histidine (H), and serine (S) along with the conserved asparagine (N) (all shown in red) of the oxy-anion hole. The amino-terminal pro-domain (purple) acts as an intra-molecular chaperone, assisting in the correct folding of the enzyme. B) Schematic organization of a plant subtilisin-like enzyme. The plant enzymes share the basic arrangement of conserved residues, but differ in several features. An amino-terminal targeting sequence (orange) is typically present. A large stretch of sequence is present between the reactive serine and the conserved asparagine. A large carboxy-terminal stretch of sequence is also commonly observed.

Subtilisin-like Protein Convertases (SPC's or PC's) have been found to activate peptides from inactive precursors (Bergeron *et al.*, 2000; Thacker and Rose, 2000). The sixth family, the pyrollysins, was described by Siezen and Leunissen (1997) as relatively heterogeneous and contained enzymes of both bacterial and plant origins. All known plant subtilisin-like proteases fall into this family. The one conspicuous feature of the plant enzymes seemed to be the presence of a large stretch of amino acids (approximately 180 aa) between the conserved asparagine and the reactive serine residue (Siezen and Leunissen, 1997), relative to the prokaryotic and other eukaryotic proteins.

This work has two primary objectives. The first is to describe the apparent heterogeneity within the pyrolysin sub-family of subtilisin-like proteases. The second is to present an approach based on the principles of comparative biology and molecular sequence analysis that will be applicable to the evolutionary analysis of large multi-gene or multi-protein families.

The plant pyrollysins have been much less intensively studied than either their bacterial or mammalian counterparts. The first somewhat characterized member of the plant subtilisin-like enzymes was Cucuminisin, a protease isolated from the melon fruit of *Cucumis melo* L. (Yamagata *et al.*, 1994). This protease showed several interesting features, particularly its production in an inactive state, and its potential activation by cleavage of its amino-terminal and carboxy-terminal peptides. The amino-terminal region of the bacterial subtilisins has been identified as an intra-molecular chaperone which aids in the correct folding of this enzyme after synthesis (Shinde and Inouye, 1994; Fu *et al.*, 2000; see Figure 8).

This has also been demonstrated in the animal PC's where the amino-terminal domain (the pro-domain) has been found to also act as an inhibitor of proteolytic activity (Thacker and Rose, 2000). The inhibition of proteolytic activity by an intramolecular sequence allows for the post-translational regulation of functional activity even when substrate specificity may be broad. The tomato SBT1 protease demonstrated a pH-dependent activation suggesting that this protease is inactive until it is secreted into the acidic apoplast (Janzik *et al.*, 2000). In *Arabidopsis*, the ARA12 protease has been found to localize to the extracellular spaces in stem tissue and was found to be optimally active under acidic conditions (pH 5) (Hamilton *et al.*, 2003). Few plant subtilases have been characterized biochemically and this will certainly need to be addressed in the future.

The identification of genes coding for plant subtilisin-like enzymes (pyrolysins) have been more numerous. The tomato *P69* gene family (Jorda *et al.*, 1999), as well as sequences from *Arabidopsis* (*AIR3*, Neuteboom *et al.*, 1999; *XSP1*, Zhao *et al.*, 2000); soybean (Batchelor *et al.*, 2000; Beilinson *et al.*, 2002); alder (Ribeiro *et al.*, 1995); and rice (Yamagata *et al.*, 2000) among others have been identified and characterized to some degree. More recently, high-throughput techniques are identifying large numbers of differentially-expressed protease genes. Three serine protease genes with homology to the *Arabidopsis* *Xylem Serine Protease1* (*XSP1*) gene were identified by a comprehensive microarray analysis of developing tracheary elements (Demura *et al.*, 2002). The specific regulation patterns observed in most of these cases suggests that there

are specific *in vivo* functions for many of these proteases. What most of these functions are is still largely unknown. The response of some of the p69 genes to pathogen exposure (Jorda *et al.*, 1999), as well as the observation that one of the few known peptide hormones in plants, systemin, is activated by a proteolytic cleavage event (Schaller and Ryan, 1995), suggests that proteolytic activity is important in plant defense. In *Arabidopsis*, two functional roles for pyrolysin subtilisin-like serine proteases have been identified. The *SDD1* gene was found to affect the distribution and patterning of stomatal density (Berger and Altmann, 2000). The *ALE1* gene was identified as regulating the normal formation of the epidermis in *Arabidopsis* embryos (Tanaka *et al.*, 2001). These findings suggest that proteolysis may also play important roles in plant development.

The large number of protease genes being identified, along with the diverse functional roles so far ascribed to some, are challenging biologists to reassess the importance of proteolysis in cellular, developmental and organismal biology. Our model systems will undoubtedly continue to provide basic understanding of the general features of organisms. The pace of large-scale sequencing efforts will generate vast amounts of data that will not be feasible to study through traditional means. Informatic approaches will be necessary to bridge the gap between the established biological and biochemical features of organisms, and the massive diversity present within the genomes and proteomes of life on earth. As mentioned, this work is an attempt to characterize the relationships between proteases of the pyrolysin sub-family of plant subtilisin-like enzymes. Through this work, a general strategy is presented that may be of

utility in the analysis of large multi-gene or multi-protein families within related organisms.

4.2. Methods

4.2.1. Sequence Data

All plant sequences were obtained from Genbank (www.ncbi.nlm.nih.gov) initially by using the BLAST algorithm and the protein sequence of At4g34980 as an initial query. This was the protease gene identified in the initial microarray experiment that was upregulated in the fiberless condition in the *Arabidopsis* hypocotyl (S. Regan, unpublished results). Additional rounds of screening were performed using hits from previous rounds. Sequences were not considered if their BLAST e-values were smaller than 10^{-20} or if they lacked the conserved residues present in this protease class (Figure 8). All sequences used in this work are identified by their gi| value and can be found in Appendix 3. Full length sequences were aligned using the ClustalW server at the European Bioinformatics Institute (www.ebi.ac.uk/clustalw/) using the default settings and further aligned manually. Alignments were viewed and manipulated using the Bioedit v.5.0.9 software. (www.mbio.ncsu.edu/BioEdit/bioedit.html)

4.2.2. Phylogenetic Reconstructions

Three methods of phylogenetic inference were used in this work. The Neighbor-Joining method was used through the Phylip package of programs. Two

Maximum-Likelihood methods were used. The program MrBayes(v.3.0, available at <http://morphbank.ebc.uu.se/mrbayes/>) was used to generate Bayesian inferences, and quartet puzzling with the program Tree-Puzzle was used. The Bayesian analysis was performed on a desktop PC while Phyllip and Tree-Puzzle were performed on the Canadian Bioinformatics Resource (CBR) server (www.cbr.nrc.ca). The parameters used for each analysis are listed in Table 1. Indels were treated as unknown amino acids and these sites therefore were not considered informative. Bootstrap analysis of the NJ inference was performed using 100 replicates as a method of assessing tree reliability. Trees were viewed and manipulated using TreeView (v.1.5.2 available at <http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>) and Treetool (v.2.0.2).

Table 1: Parameters used for phylogenetic analysis

Phylogenetic Method	Substitution Model	Rate Heterogeneity?	Misc.
Neighbor-Joining	Kimura	No	
Bayesian	PAM	Yes, approximated with a discrete gamma distribution, 4 categories	MCMC analysis: 200000 generations, sampled every 100 generations; 4 chains; burnin=50000 generations
Quartet Puzzling (ML)	WAG	Yes, approximated with a discrete gamma distribution, 8 categories	Parameter estimation was approximate

4.2.3. Structure Prediction

Hypothetical protein structures for the *Arabidopsis* subtilisin-like protein family were generated using the SwissModel protein modeling server (<http://swissmodel.expasy.org/>). Templates were selected using the default settings using a BLAST value of 10^{-5} . Structures were viewed and manipulated using the Swiss pdb viewer (v.3.7 available at <http://www.expasy.org.spdbv/>). Molecular surfaces and electrostatic potentials were calculated using the default parameters. Two-dimensional images were manipulated using Adobe Photoshop 5.0 and Adobe Illustrator 8.0.

4.3. Results and Discussion

4.3.1. The Plant Subtilisin-Like Proteases

The protein sequences used in this work are listed in Appendix 3 along with their organism of origin and presumed function. Those without reference information were direct submissions by their authors to Genbank. The multiple sequence alignment of 102 plant subtilisin-like proteases (Appendix 4) was used to generate trees describing the evolutionary sequence relationships of these protease sequences. Approximately half of these sequences (55) are from *Arabidopsis*, approximately one quarter (25) are from rice and the remainder are from various plant species. With the exception of one, the AF70 from spruce, all are from angiosperms. The Neighbor-Joining (NJ) and Bayesian (B) inferences are shown (Figures 9-11). The general topologies of these trees agree with each other and the support values at the various nodes are given (Figure 9). Large sequence groups that were consistently derived are (somewhat arbitrarily) identified by numbers 1-8. This was done to aid in later analyses. The biological relevance of these groups will be discussed later. The first value given at each node is the posterior probability of the group as determined by the Bayesian analysis. The second value is the bootstrap statistic of support based on 100 replicates. When the topologies from these two analyses did not agree, only the posterior probabilities are given. The posterior probabilities are generally higher than the bootstrap values. This may be a result of an observed overestimation of support that has previously been reported in Bayesian analyses (Huelsenbeck *et al.*, 2002). This may also be due to the different models of amino acid

Figure 9: Sequence relationships of the plant subtilisin-like sequences. The topology is based on the Bayesian inference of phylogeny. The values at the nodes are the support values for the relationships indicated. The first value is the posterior probability as estimated by the Bayesian analysis. The second value is the bootstrap statistic of 100 replicates of the Neighbor-Joining analysis. An asterisk indicates that the topology was not shared between these two analyses.

Figure 10: Unrooted radial tree of plant-pyrolysin sequence relationships based on Neighbor-Joining analysis. Eight groups of sequences were defined based on observed clustering tendency.

0.10

Figure 11: Radial unrooted tree of plant pyrolysin sequences based on a Bayesian inference of sequence relationships. The eight groups recognized in the Neighbor-Joining analysis are also identifiable in this analysis.

substitution used during the two analyses. These two explanations are not mutually exclusive and could both be contributing to the observed discrepancy. Despite these potential problems, I believe that the common topologies observed in both methods of reconstruction and the generally good levels of support observed indicate that the sequence relationships observed are genuine.

The topologies of gene trees i.e. those including all homologous sequences, are usually presented as unrooted, radial trees. Presumably this is because traditional phylograms (Figure 9) impose an outgroup/ingroup structure on trees when there is usually little basis for identifying which sequence best represents the ancestral state. This is evidenced in Figure 9 where ARA12 is presented as the outgroup. To avoid implying that any one sequence is a better outgroup than any other in this analysis, trees will generally be presented as unrooted radial trees (e.g. Figure 10). When viewed as unrooted trees, the sequence relationships relative to each other are more obvious. The position of ARA12 is still within group 4 but the arbitrary choice for it as an outgroup is apparent. Several other features of these sequences are easily visualized on these trees.

Group 1 sequences form the most distinctive group on these trees. This is due to the long branch which separates this group from all others and indicates a large degree of shared sequence similarity of group 1 members relative to the other sequences included in this analysis. The basis for this shared similarity is not known. The most likely explanations for this shared similarity are descent from a common ancestral sequence, sequence convergence due to selection for

similar function, or possibly some combination of the two. Which of these is the case is an extension of the classical problem of assessing homologous versus homoplasious relationships of characters in evolution. This cannot be adequately assessed from the present analysis except in a few instances that will be described. The classical approach to this problem is to use the comparative method to characterize specific features of the two characters in question. Several members of group 1 have been partially characterized. The *LIM9* gene of *Lillium longiflorum* was identified and found to be differentially expressed during microsporogenesis (Taylor *et al.*, 1997). The highly similar tomato Tmp gene was found to be similarly expressed (Meichtry *et al.*, 1999). Detailed functional characterizations for both of these are lacking however their high sequence similarity along with their similar patterns of expression make it tempting to speculate that they perform similar roles *in vivo* and are orthologs. While none of the results of the present work directly assess this prediction, there are several features that can be addressed. If LIM9 and TMP are orthologs, it is curious that there are no *Arabidopsis* or rice orthologs that more obviously group with these two sequences. There are several plausible explanations for this. The search for rice sequences in the databases was not exhaustive, therefore a rice ortholog may exist that was not included in this analysis. The retrieval of *Arabidopsis* sequences is believed to have been more thorough and it seems unlikely that unintentional omission would be the case for *Arabidopsis*. The most similar *Arabidopsis* sequence to LIM9 and TMP is the ALE1 protein. This protein has been implicated in regulating the differentiation of the epidermis in

Arabidopsis embryos (Tanaka *et al.*, 2001). It is not readily apparent why this protease should be the most similar to LIM9 and TMP. The long branch of ALE1 to the LIM9/TMP group should be noted and implies significant differences between these sequences.

Gene loss is a potential explanation for this apparent lack of orthology. The tomato Sbt4d sequence (not included in this analysis) was found to be interrupted by the insertion of a gypsy-like retrotransposon (Meichtry *et al.*, 1999). In the course of this work it was observed that the rice sequences gi|22773236 and gi|22773238 were complementary truncated sequences i.e. they would constitute a single complete protease sequence if joined. The intervening sequence, gi|22773237, is annotated as a Miniature Inverted Transposable Element (MITE). It is highly likely that at some point in the past this was an intact sequence that was interrupted by the insertion of this element. Other truncated sequences were noted during the course of this work but it is not known how these were biologically generated. This analysis again cannot make predictions regarding specific protein functionality i.e., whether disrupted or truncated sequences retain an *in planta* function. Despite this, it seems likely that most disrupted sequences become pseudogenes. That this is not universally true may gain credence from the origins of the plant subtilisin-like proteases. As stated earlier, the plant pyrollysins are characterized by a large, 180 aa sequence stretches between the catalytic serine and the asparagine of the oxy-anion hole (Siezen and Leunissen, 1997), relative to most prokaryotic and all known eukaryotic members of the subtilisin family. The presence of this

region suggests that some small insertions could be tolerated. This would likely depend on whether the protease could maintain the structural integrity of the active site and thereby maintain functional activity.

Gene loss may not be the only explanation for the lack of *Arabidopsis* sequences closely associated with TMP and LIM9. Other evolutionary forces may be affecting the amino acid sequences. For example, another group 1 *Arabidopsis* sequence could possibly be orthologous to TMP and LIM9. The changes in amino acid sequence leading to the relationships observed in this group could have arisen through random genetic drift. This could occur if there were frequent changes in regions of the coding sequence not necessary for *specific* functional activity. Similar function could still be maintained if selection was acting primarily in the maintenance of regulation or expression. Selective constraints may keep the expression of this gene coordinated to a specific time and developmental stage and, as long as this pattern of expression is maintained, any broad-specificity protease could act on the necessary target. In this scenario, few functional constraints would be maintained within the coding region beyond those required for basic proteolytic activity. This would result in a loss of species-level phylogenetic signal after a period of time sufficient for multiple substitutions to occur within the protein coding sequence. As mentioned above there is a large degree of difference between the group 1 members as a whole and those of the other groups, marked by the single long branch joining this group to the central portion of this tree. If this single branch represents genuine shared characteristics of the amino acid sequences of this

group (and not just an example of long-branch attraction) it will be interesting to see what common features these proteases have distinct from the others included in this analysis. This will not be possible until detailed biochemical and functional characterizations of many of these proteases have been performed. The preceding discussion highlights some of the difficulties associated with hypotheses of orthology within large protein families. These difficulties need to be addressed not only within individual sequence groups but across them as well when so little characterization of these proteins has been done.

One feature commonly observed in many of these sequence groups is the presence of subgroups of very closely related sequences. The P69 group has previously been identified as a tandem repeat of protease genes in tomato (Jorda, 1999). The physical structure of these genes represent good evidence concerning the precise evolutionary history between these highly similar sequences. Tandem repeats of the genes coding for these proteases are observed in several groups upon inspection of the genomic regions that code for these proteases. Large clusters containing four or more sequences are found in group 2 (rice), group 3 (tomato P69's), group 5 (*Arabidopsis*), and group 7 (*Arabidopsis*). Several smaller tandem repeats are observed in this protein family in both *Arabidopsis* and rice. The abundance of these suggests that this is a relatively common event over evolutionary time. The relatively well characterized example of the tomato P69 genes allows us to examine the functional roles of these highly similar sequences. The P69 protein sequences are highly similar to each other (Meichtry *et al.*, 1999). Among the 6 family

members the lowest level of amino acid similarity exists between P69F and P69C which are 79% similar. The highest similarity exists between P69A and D which are 98% similar. Despite these similarities, it has been demonstrated that these proteases are differentially regulated at the transcriptional level (Jorda, 1999). *P69A* and *D* were found to be constitutively expressed but at slightly different levels in by RT-PCR and Southern-blot analysis. *P69B* and *C* were found to only be expressed during exposure to the pathogenic bacteria *Pseudomonas syringae*. Reporter-gene fusions to the various *P69* promoter regions also showed varying patterns of expression in transgenic *Arabidopsis* plants. The *P69E* and *F* genes were also found to be differentially regulated showing highly localized expression in roots and hyathodes respectively (Jorda *et al.*, 2000). These patterns of differential expression led those authors to suggest that the different *P69* proteins fulfills different roles *in vivo*. If this is true then it would seem that the evolution of gene regulation can be uncoupled from the evolution of the protein sequence and that each can independently contribute to the fitness of the organism. Whether this is the case has not been conclusively demonstrated but seems entirely possible.

The limits of the current analysis in predicting the functional roles of proteases based solely on sequence similarity are generally understood. That this limitation is in part due to limits in the power of this method is evident. Despite this, I believe however that the power of these methods have not been fully exploited. Our knowledge of the organisms containing these proteases is another level of biological organization that is well documented. It therefore

seems reasonable to individually examine the protease families of several plant species whose sequences were included in the previous analysis. Sequences derived from *Arabidopsis*, rice and tomato were extracted from the original alignment and collected separately to generate the pyrolysin families from these three species (Figures 12-14). These trees were oriented so that sequence groups were located in the same relative positions as previously presented for ease of comparison.

The analysis initially identified 55 members of the pyrolysin class of subtilisin-like proteases in *Arabidopsis*. Another member of this class was independently identified and was excluded from the comparative analysis for reasons that will be discussed later. The NJ and ML trees show the sequence relationships of the *Arabidopsis* pyrolysins (Figure 12). The general groupings of sequences agree with those seen in the previous analysis. Some of the basal relationships in the ML tree are poorly resolved. This inference method joins several groups at single nodes that are poorly supported. Here we see that the basal branching order of groups 2, 3, 4, and 5 are not resolved. While this is generally non-informative in species level phylogenetics, it seems less important in the analysis of molecular sequences. The important feature of these trees is that the distinctness of the individual sequence groups is maintained. One instance where this is not observed is with the group 8 sequence At5g67090. The NJ analysis maintains this sequence as distinct, branching from the group 1 branch at the base of the tree. On the ML tree this sequence is placed as a member of group 1. This poor resolution of this sequence is also evident when

Figure 12A: The Arabidopsis pyrolysin family. Inference of sequence relationships based on Neighbor-Joining analysis. The numbers in brackets correspond to the identifiers used in the structural analysis. All groups identified previously are observed with the exception of Group 7.

Figure 12B: The Arabidopsis pyrolysin family. Inference of sequence relationships based on Maximum-Likelihood analysis. Note the poor resolution of the basal relationships in the central region.

Figure 13A: The rice pyrolysin family. Sequence relationships based on NJ analysis. All sequence groups are represented except for Group 5.

Figure 13B: The rice family of pyrolysin subtilisin-like proteases as inferred by ML. Note the poor resolution of the central region of this tree.

Figure 14: The tomato pyrolysin family as inferred by NJ analysis. Four sequence groups are present, Groups 1, 6, 7, and 8.

observed on the Bayesian tree that shows At5g67090 within group 2. Although this sequence was anomalous, most sequences maintained their grouping when observed within the species delimited trees.

25 sequences were obtained in the non-exhaustive search of the available *Oryza sativa* sequences. Again a single sequence was excluded. The sequence relationships by NJ and ML (Figure 13) again show sequence groupings similar to those in the larger plant sequence analysis (Figures 9-11). Some sequence groups are not represented within the rice family presented here. Whether this represents the incomplete search of the rice genome or the absence of these sequences within rice is not known. No rice sequences were observed for group 5. Group 2 in the rice ML tree showed poor resolution at its base. This is not surprising considering the highly diverse character of this group (Figure 13B). The difference in branch lengths observed for the rice sequence gi|21740608 between the two methods is probably due to the different assumptions made between the two methods. This sequence appeared to be truncated at the amino terminus and was only roughly one-third the size of the other protease sequences. The tomato family had been previously analyzed (Meichtry, 1999) and the results here based on NJ analysis (Figure 14) agree with those previous findings. Four of the sequence groups identified in this work are represented by the known tomato sequences. Members of groups 1, 6, 7, and 8 are present. The limited extent to which the tomato genome has been characterized leaves open the possibility that more subtilisin-like proteases may

exist in tomato. This possibility was acknowledged during the previous characterization of this protease family (Meichtry *et al.*, 1999).

The presence of members of the same sequence groups within different species makes it tempting to speculate about the nature of this protease family in evolutionary history. Caution must again be used due to the problems discussed previously regarding the evolution of sequences with respect to phylogeny (origin) versus evolved function. The possibility of a large number of sequences not included in the tomato family presented here makes it a poor candidate for inclusion in this strategy. As an initial approach, each sequence group identified could be represented by one sequence in the common ancestor of *Arabidopsis* and Rice, an organism that predates the monocot/dicot divergence. This hypothetical ancestral protein family (Figure 15) has branches representing progenitors of groups 1, 2, 3, 4, 6, and 8. The basal branching relationships are shown as unresolved due to the poor resolution of these among extant sequences (Figures 4 and 5). This is probably a most naïve approach to qualitatively modeling the ancestral state of a protein family. The limited sequence data, along with the large time frame between divergence of monocots and dicots both introduce potentially erroneous assumptions implicit in the model presented (Figure 15). This approach may have much greater utility (at least initially) if we consider two much more closely related species. For example, if we were to consider another member of the *Brassicaceae* and compare it's protease family to that of *Arabidopsis*, what would we expect? Our understanding of evolution would predict that fewer changes would have taken

Figure 15: Hypothetical ancestral state of the plant subtilisin-like protease family prior to the monocot-dicot divergence. The groups present are those that are present in both of the extant species used to infer this family.

place between this organism and *Arabidopsis* than between rice and *Arabidopsis* as they share a much more recent common ancestor. This close a comparison could help not only to construct more accurate models of protein family evolution but could potentially highlight important developmental or regulatory differences between closely related species through the tracking of specific gene loss or gene duplication events.

4.3.2. Assessing the Properties of Sequence Groups

The evolutionary significance of groups of similar sequences is central to this work. One potentially useful approach to assess this would be to describe the physicochemical properties of sequences and observe potential correlations of these properties within and across groups. In other words, to extend the comparative method to the realm of molecular sequence evolution. Properties such as structural information, enzymatic activity, substrate specificity, charge distribution etc. could be potentially examined within an evolutionary framework. Many of these properties can only currently be adequately examined through traditional biochemical means. The immense amount of sequence data presently being generated provides avenues of investigation previously unavailable. I have chosen to undertake an informatics-based approach to structurally characterize the *Arabidopsis* subtilisin-like protease family and to observe whether the physical properties predicted by protein modeling are consistent with the evolutionary framework of these sequences.

Protein structure modeling by threading analysis can be used to predict three-dimensional protein structure (Madej *et al.*, 1995). This technique uses known three-dimensional structural information of empirically derived models to predict the structure of previously unknown proteins. A protein-modeling server is publicly available for structural prediction studies (Guex and Peitsch, 1997). Query amino acid sequences are “threaded” through structures contained in the database. The “threading” consists of the evaluation of how well the query sequence can adopt the same three-dimensional structure as the template based on the energetic properties of its amino acid sequence. Each of the members of the *Arabidopsis* subtilisin-like family that was included in the phylogenetic analysis was subjected to threading analysis. The templates selected by the server for comparison to each sequence are given in Appendix 5. A large amount of structural information is available for the bacterial subtilisin-like enzymes due to intensive study centering around their use in industrial applications. No plant subtilisin-like proteases have been structurally characterized by X-ray crystallography or Nuclear Magnetic Resonance (NMR) Spectroscopy to date. All of the templates selected are structures of bacterial enzymes. This may introduce a source of error as it is not known how the 3-D structure of the plant proteases might differ from those of the bacterial enzymes. As previously mentioned the domain structure of the plant enzymes differs from most bacterial subtilases (Figure 16A). Plant subtilisin-like proteases often have i) targeting sequences at the amino terminal for direction to the endomembrane system, ii) large stretches of amino-acid sequence between the reactive serine

Figure 16: Structural modeling of the plant subtilisin-like proteases. A) Schematic view of the minimum consistently modeled region, shown in yellow on the bacterial protease 1MPT (top) and in blue on the plant sequence modeled, the *Arabidopsis* sequence At1g20150 (bottom). This region encompasses the reactive serine residue in both proteins. B) Structural modes of the two proteins in A. The bacterial structure (left) was previously determined by X-ray crystallography while the plant structure (right) was predicted by threading analysis, using 1MPT as a template. The region modeled is shown in yellow in the bacterial enzyme and in blue in the plant structure. Active-site and conserved residues are red.

B

and the conserved asparagine, as well as iii) often having carboxy-terminal pro-peptide sequences (in addition to the amino-terminal pro-peptides) all of which seem to be absent in the well characterized bacterial subtilisins.

The results of the modeling suggest that only a limited portion of the plant proteins share structural similarity to the bacterial proteases (Figure 16B). The only region that was consistently modeled was an approximately 60 amino acid region surrounding the serine of the reactive site. A few plant sequences showed more extensive structural similarity while others could not be modeled at all. This active-site serine residue is found at one end of an alpha helix in both the template bacterial structures and the derived hypothetical plant structures. In the bacterial enzymes, the folding of the protein allows for all the active site residues to be brought into an appropriate, spatially close orientation, in order to achieve catalytic activity (Creighton, 1993). From these results it cannot be assessed how the three dimensional active-site structure of the plant proteases relates to that of the bacterial enzymes.

Despite the relatively poor extent to which the plant sequences were structurally modeled, other, structurally-based physicochemical properties of these enzymes could still be evolutionarily tractable. The Swiss pdb viewer program allows for the computation of molecular surfaces based on structural models. It also allows for the computation of the electrostatic potential of these surfaces. These were calculated for each predicted structure generated. Each of these structures was then mapped onto the NJ sequence phylogeny (Figure 12A) to see if structural or charge distribution patterns, coinciding with raw

sequence relationships were evident (Figure 17). In some instances, particularly when branches describing sequence relationships are long (and therefore similarity low) few obvious patterns beyond basic geometry are observed. In contrast to this, when examining more similar sequences, more obvious patterns are observed. For example, structures 12, 13, and 14, representing sequences At5g59100, At3g46850, and At3g46840 respectively, share both geometrical similarities as well as patterns of charge distribution. This is particularly interesting because these latter two are a tandem cluster in the genome, while At5g59100 is coded for by a gene on a different chromosome. Additionally, the At5g59100 gene is part of a tandem array with three other sequences on chromosome 5. These are represented by structures 5, 6, and 7, a seemingly structurally more diverse group. These kinds of data could best be applied in a comparative framework with closely related species to delineate specific evolutionary relationships. A potential pitfall to these kinds of comparisons would be to disregard the potential affects of post-translational modifications. One example of this could be glycosylation, as many of the plant subtilisin-like proteases have been shown to be linked to various sugars (Yamagata *et al.*, 1994; Taylor *et al.*, 1997; Bogacheva, 1999). Our understanding of the biological roles of glycosylation is far from complete (Wilson, 2002) but it has been demonstrated that glycosylation can affect the enzymatic properties of proteins (Murphy and Kirley, 2003). One alternative approach, perhaps more practical while the amount of available sequence data is relatively small, would be to look for sequence similarities in non-coding regions near these gene tandems or at

Figure 17: The sequence relationships of the *Arabidopsis* subtilisin-like proteases can be simultaneously traced by physicochemical features. Hypothetical protein structures are mapped onto the NJ inference of *Arabidopsis* sequence relationships. Blue regions represent regions of positive electrostatic potential, red regions correspond to regions of negative potential. White regions are neutral. Numbers correspond to those on Figure 12A.

intron/exon structure. These would also strongly support hypotheses of homology as it is presumed that these regions would be free to evolve without severe selective constraints. Any similarity thus observed should be due to shared ancestry.

Although it seems reasonable that many features of these proteases could be evolutionarily tractable, the scheme presented here may suffer from an inherent failure to account for genetic processes that are not characterized by direct vertical descent. Amongst micro-organisms, lateral (horizontal) gene transfer is viewed as an important evolutionary mechanism. Until recently, this has been viewed as important chiefly among prokaryotes but recent work shows it to be important in microbial eukaryote evolution (Anderson *et al.*, 2003). This may have been more important in the early history of this protein family within the plant lineage as lateral transfer is not thought to be a significant factor in the evolution of large multicellular eukaryotes (Salzberg *et al.*, 2001). Another process that may be more important in recent evolutionary history is the possibility of recombination between members of this family. The extent to which this would affect this scheme is not known. This would probably depend on the frequency of recombination between any two family members as well as how recombined sequences were acted upon by selection. The extent of these factors is not known and would likely need to be experimentally determined.

4.3.3. The Origins of Plant Pyrollysins

The plant subtilisin-like proteases have been placed in the family of the pyrollysins by Siezen and Leunissen (1997). This group was described as generally heterogeneous in terms of sequence relationships and contains sequences derived from various groups of bacteria as well as plants, the only eukaryotic taxon previously believed to contain pyrollysins. The present analysis has been an attempt to further describe this apparent heterogeneity within plants. In an attempt to characterize the sequence relationships between the bacterial and plant pyrollysins, several rounds of BLAST searching were performed by querying plant sequences to database searches limited to prokaryotic sequences. The absence of fungal and metazoan pyrollysins was noted by Siezen and Leunissen (1997). Since their work, a tremendous amount of molecular sequence data has been generated, principally by the large-scale genomic sequencing projects. Literature and sequence database searches were performed to see whether pyrollysins from other eukaryotic groups had been recently identified. The literature search identified several mammalian sequences (denoted SKI1 or Site1) that had since been characterized as pyrollysins (Touré *et al.*, 2000). All of these sequences appear to orthologous and have been shown to be involved in the cleavage of sterol regulatory element binding proteins (SREBP's), regulating the pathway of cholesterol and fatty acid metabolism. The sequences for these enzymes were obtained from Genbank and used to search for similar plant and bacterial sequences. Two plant sequences were identified, At5g19660 and a rice sequence denoted here as

Rice.1. These were the sequences mentioned that were omitted from the previous analysis.

The sequences from both searches were grouped together and subjected to ML analysis to determine the sequence relationships of various animal, plant and bacterial pyrolysins (Figure 18). Several features of interest are observed in this tree. The first, and perhaps most obvious, is that the central portion of this tree is very poorly resolved. Most of the bacterial sequences are separated by long branches from the central node (N). Some of the plant sequences from the previous analysis were included and these show a strong grouping with each other, distinct from most of the bacterial sequences included. There were some interesting exceptions to this. Sequences from *Shewanella* and *Chloroflexus* consistently grouped with the plant sequences. *Shewanella* belongs to the gamma-proteobacteria while *Chloroflexus* is a member of the green non-sulfur or green filamentous bacteria (Pierson and Castenholz, 1974). The previous analysis that grouped only plant and bacterial sequences within the pyrolysin group led Bogacheva (1999) to hypothesize that the plant subtilisins could be derived from a protease harbored by the endosymbiotic ancestor of plastids and that all plant subtilisins have been derived from this ancestral state. Over evolutionary time, the gene for this protease was transferred to the nucleus and continued to evolve into the diverse family of proteases currently observed. This is in line with recent findings concerning the large number of gene transfers believed to have taken place from the endosymbiont within plants (Martin *et al.*, 2002). The results presented here are consistent with the hypothesis that the

Figure 18: Maximum Likelihood estimate of sequence relationships of various members of the pyrolysin sub-family of subtilisin-like serine proteases. Numbers indicate the likelihood values supporting the observed relationships. The central node (N), plant sequences, and the Site1 group are the prominent features.

_0.1

plant pyrolysins analyzed here (Figure 9) were derived from a prokaryotic endosymbiont. The findings here do show several curious features of sequence relationships. *Chloroflexus* is consistently found to be the most similar prokaryotic sequence but it is generally believed that chloroplast is more closely related to the cyanobacteria than the green non-sulfur bacteria (Xiong *et al.*, 2000; Martin *et al.*, 2002). The evolutionary relationships among these groups has been difficult to discern conclusively (Gupta *et al.*, 1999; Xiong *et al.*, 2000) and our understanding of the nature of the endosymbiont may be lacking. The complete genome sequence of two cyanobacteria are known and subtilisin-like sequences were included in this analysis from *Synechocystis* and *Nostoc* (Anabaena). None of these cyanobacterial sequences showed close relationships to the plant sequences. There are several plausible explanations for this. It is possible that gene losses occurred in the cyanobacterial species included in this analysis. Ancient lateral transfers could also be interfering with these inferred relationships. Even though the data presented here might suggest these possibilities, care must be taken when inferring very ancient relationships. Another possibility that could potentially explain this result is one based on methodology. The sequences used in this study were full-length sequences and included the large insertion characteristically found in the plant sequences. The *Shewanella* and *Chloroflexus* sequences also possessed a large sequence insertion at the same location as the plant sequences. If these insertions were generated independently, the presence of this in the analysis may have led to an overestimation of the relatedness between these prokaryotic sequences with the

plant sequences. A more comprehensive multi-locus analysis would better assess predictions regarding ancient relationships.

The animal pyrolysins that were previously mentioned showed no obvious relationship to any of the plant or bacterial sequences except for the two plant sequences mentioned. The relationships within this group show a very high level of similarity. The relationships derived by ML analysis also show a very clear species phylogeny amongst these. The plant sequences group strongly together, distinct from the animal sequences. The two invertebrate sequences also group together as do the mammalian sequences. These findings strongly suggest that these genes are orthologous and that they have evolved within the eukaryotic lineage since before the plant-metazoan divergence. One caveat to this conclusion however is that no *Caenorhabditis* or fungal homologs of SKI1 were found. Despite the potential challenges to these relationships, the emerging evolutionary picture of plant pyrolysin sequences is that the vast majority of them are much more similar to the bacterial sequences analyzed here. This is consistent with Bogacheva's hypothesis of endosymbiotic origin for these enzymes. Another group of pyrolysins, represented by one sequence in *Arabidopsis* and one in rice and bearing very high similarity to the mammalian SKI1 protease probably resided in the ancient host genome that was the common ancestor of plants and animals.

4.3.4. Towards a global model of sequence evolution

Models of sequence evolution have generally taken a bottom up approach. That is, molecular evolution has been primarily concerned with deriving the general statistical probabilities of micro-evolutionary events, and then using these to assess particular sequence relationships. These methods have been tremendously successful in assisting the evolutionary biologist in clarifying relationships that otherwise may not have been possible. The models have been increasing in complexity in recent years and are beginning to accommodate more and more biologically relevant information (Whelan *et al.*, 2001). Can macro-evolutionary data be integrated into molecular evolutionary models? If our understanding of the living world as a continuum of descent with modification is to become a workable whole, then at some point macro- and micro-evolutionary data must merge.

This work has attempted to characterize the sequence relationships of the plant subtilisin-like proteases. In doing so, some of the evolutionary relationships between these sequences seem quite tractable while others much less so. An attempt has been made to use species-level information to bring another level of biological organization within the sequence evolution framework. As a conceptual model, each extant species' pyrolysin family can be visualized as existing on the surface of a sphere (Figure 19). Groups of taxonomically related species could be grouped together on the surface of the sphere in taxonomically meaningful ways. For example, monocots distinct from dicots, angiosperms distinct from gymnosperms etc. If one were to peer beyond the surface of this

Figure 19: The evolution of multi-protein families can be modeled through a spherical geometry. Extant species are projected onto the surface of this sphere. Inferences based on these can be used to predict the ancestral paths of family and sequence evolution through to the centre of this structure: the hypothetical ancestor of all homologous sequences.

sphere, within the interior would lie the paths traced by the organismal lineages i.e. the paths traced by traditional phylogenetics. Instead of being represented by single lines tracing the organismal path, we would observe the two-dimensional framework of sequence-family relationships inferred from those present within extant species (these could be visualized as nodes e.g. Figure 15). The ancestral sequence that was the common ancestor to all extant sequences would lie at the centre of this sphere. The utility of such a conceptual model might be the intuitive picture of sequence evolution gained when generated within a computer-based environment. As a database structure, this kind of model could allow for the storage and presentation of vast amounts of evolutionarily tractable data. The use of object-oriented programming would seem particularly suited to such a task.

5. General conclusions

The molecular genetics of wood formation are only beginning to be uncovered. Modern high-throughput techniques have the capacity to identify large number of genes and their products involved in this highly regulated process. This identification is only a first step in understanding the developmental and evolutionary features of xylem formation. Detailed biochemical and cellular studies will use this information to build a more complete description of these processes. This work is the characterization of two plant protein families whose members may be involved in wood formation. They represent two examples of potentially different levels of control. The first family considered, the ethylene

receptors, are coded for by a relatively small gene family in most species investigated. They govern the perception of a message carried by one of the classical plant hormones. The large amount of data concerning the physiological roles of ethylene in plants has not been synthesized into a cohesive framework in light of the more recent developments concerning the perception of this stimulus. While the link between ethylene perception and effect has not been elucidated, the effects will likely be governed by second messenger systems due to their apparent diversity. The second family studied in this work could be responsible in part for this diversity. The *Arabidopsis* subtilisin-like protease family has 56 members as identified in this work. The limited characterization of the physiological roles of these proteases suggests that they may govern an even wider range of roles than that of the ethylene receptors. They could be highly regulated, both at the genic as well as protein activity levels and could be acting as either downstream signalling regulators or ultimate effectors.

The work presented here has attempted to characterize the diversity of protein family members and investigate specific roles in a simpler case i.e. the ethylene receptors. In a more complex case, the large number of proteases studied combined with limited characterization of the various members makes any functional generalizations premature. Nevertheless, the large amount of sequence data likely to be generated in the near future will necessitate an improved informatic approach consistent with the observed features of biological entities.

6. References

- Abeles, F.B., Morgan, P.W., and Saltveit, M.E. (1992) Ethylene in Plant Biology. 2nd edition. Academic Press, San Diego, CA, U.S.A.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
- Anderson, J.O., Sjögren, Å.M., Davis, L.A.M., Embley, T.M., and Roger, A.J. (2003) Phylogenetic analysis of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Curr. Biol.* 13: 94-104.
- Batchelor, A.K., Boutilier, K., Miller, S.S., Labbe, H., Bowman, L., Hu, M., Johnson, D.A., Gijzen, M., and Miki, B.L.A. (2000) The seed coat-specific expression of a subtilisin-like gene *SCS1*, from soybean. *Planta* 211: 484-492.
- Beilinson, V., Moskalenko, O.V., Livingstone III, D.S., Reverdatto, S.V., Jung, R., and Nielsen, N.C. (2002) Two subtilisin-like proteases from soybean. *Physiol. Plant.* 115: 585-597.
- Berger, D., and Altmann, T. (2000) A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev.* 14: 1119-1131.
- Bergeron, F., Leduc, R., and Day, R. (2000) Subtilase-like pro-protein conertases: from molecular specificity to therapeutic applications. *J. Mol. Endocrinol.* 24: 1-22.

- Bleecker, A.B., Estelle, M.A., Somerville, C., and Kende, H. (1988) Insensitivity to ethylene conferred by a dominant mutation in *Arabidopsis thaliana*. *Science* 241: 269-274.
- Bleecker, A.B. (1999) Ethylene perception and signaling: an evolutionary perspective. *Trends Plant Sci.* 4: 269-274.
- Bogacheva, A. M. (1999) Plant Subtilisins. *Biochem. (Moscow)* 64: 287-293.
- Chaffey, N., Cholewa, E., Regan, S., and Sundberg, B. (2002) Secondary xylem in *Arabidopsis*: a model for wood formation. *Physiol. Plant.* 114: 594-600.
- Chang, C., Kwok, S.F., Bleecker, A.B., and Meyerowitz, E.M. (1993) *Arabidopsis* ethylene-response gene ETR1: similarity of product to two-component regulators. *Science* 262: 539-544.
- Chang, C., and Meyerowitz, E.M. (1995) The ethylene hormone response in *Arabidopsis*: a eukaryotic two-component signaling system. *Proc. Nat. Acad. Sci. U.S.A.* 92: 4129-4133.
- Chen, Y., Randlett, M.D., Findell, J.L., and Schaller, G.E. (2002) Localization of the ethylene receptor ETR1 to endoplasmic reticulum of *Arabidopsis*. *J.Biol. Chem.* 277: 19861-19866.
- Creighton, T.E. (1994) *Proteins: structures and molecular properties* 2nd edition. W. H. Freeman and Company, New York.
- Demura, T., Tashiro, G., Horiguchi, G., Kishimoto, N., Kubo, M., Matsuoka, N., Minami, A., Nagata-Hiwatashi, M., Nakamura, K., Okamura, Y., Sassa, N., Suzuki, S., Yazaki, J., Kikuchi, S., and Fukuda, H. (2002) Visualization by comprehensive microarray analysis of gene expression programs during

transdifferentiation of mesophyll cells into xylem cells. *Proc. Nat. Acad. Sci. U.S.A.* 99: 15794-15799.

Feild, T.S., Lee, D.W., and Holbrook, N.M. (2001) Why leaves turn red in autumn. The role of anthocyanins in senescing leaves of Red-Osier Dogwood. *Plant Physiol.* 127: 566-574.

Fu, X., Inouye, M., and Shinde, U. (2000) Folding pathway mediated by an intramolecular chaperone. *J. Biol. Chem.* 275: 16871-16878.

Gu, J., Wang, Y., and Gu, X. (2002) Evolutionary Analysis for functional divergence of jak protein kinase domains and tissue-specific genes. *J. Mol. Evol.* 54: 725-733.

Gupta, R.H., Mukhtar, T., and Sinhg, B. (1999) Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. *Mol. Microbiol.* 32: 893-906.

Hall, M.A., Moshkov, I.E., Novikova, G.V., Mur, L.A., and Smith, A.R. (2001) Ethylene signal perception and transduction: multiple paradigms? *Biol. Rev. Camb. Philos. Soc.* 76: 103-128.

Hamilton, J.M.U., Simpson, D.J., Hyman, S.C., Ndimba, B.K., and Slabas, A.R. (2003) Ara12 subtilisin-like protease from *Arabidopsis thaliana*: purification, substrate specificity and tissue localization. *Biochem. J.* 370: 57-67.

Hua, J., Chang, C., Sun, Q., and Meyerowitz, E.M. (1995) Ethylene insensitivity conferred by *Arabidopsis ERS* gene. *Science* 269: 1712-1714.

Hua, J., and Meyerowitz, E.M. (1998) Ethylene responses are negatively regulated by a receptor gene family in *Arabidopsis thaliana*. *Cell* 94: 261-271.

Hua, J., Sakai, H., Nourizadeh, S., Chen, Q., Bleeker, A.B., Ecker, J.R., and Meyerowitz, E.M (1998) *EIN4* and *ERS2* are members of the putative ethylene receptor gene family in *Arabidopsis*. *Plant Cell* 10: 1321-1332.

Huelsenbeck, J.P., Larget, B., Miller, R.E., and Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51: 673-688.

Janzik, I., Macheroux, P., Amrhein, N., and Schaller, A. (2000) *LeSBT1*, a subtilase from tomato plants. *J. Biol. Chem.* 275: 5193-5199.

Jorda, L., Coego, A., Conejero, V., and Vera, P. (1999) A genomic cluster containing four differentially regulated subtilisin-like processing protease genes is in tomato plants. *J. Biol. Chem.* 274: 2360-2365.

Jorda, L., Conejero, V., and Vera, P. (2000) Characterization of *P69E* and *P69F*, two differentially regulated genes encoding new members of the subtilisin-like proteinase family from tomato plants. *Plant Physiol.* 122: 67-73.

Kieber, J., and Ecker, J.R. (1993) Ethylene gas: it's not just for ripening any more! *Trends Genet.* 9: 356-362.

Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* 17: 1244-1245.

Madej, T., Boguski, M.S., and Bryant, S.H. (1995) Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Lett.* 373: 13-18.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, S., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002) Evolutionary analysis of

Arabidopsis, cyanobacterial and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Nat. Acad. Sci. U.S.A. 99: 12246-12251.

Meichtry, J., Amrhein, N., and Schaller, A. (1999) Characterization of the subtilase gene family in tomato (*Lycopersicon esculentum* Mill.). Plant Mol. Biol. 39: 749-760.

Murphy, D.M, and Kirley, T.L. (2003) Asparagine 81, an invariant glycosylation site near apyrase conserved region 1, is essential for full enzymatic activity of ecto-nucleoside triphosphate diphosphohydrolase 3. Arch. Biochem. Biophys. 413: 107-115.

Neuteboom, L.W., Ng, J.M.Y., Kuyper, M., Clijdesdale, O.R., Hooykaas, P.J.J. and vander Zaal, B.J. (1999) Isolation and characterization of cDNA clones corresponding with mRNAs that accumulate during auxin-induced lateral root formation. Plant Mol. Biol. 39: 273-287.

Ota, I.M., and Varshavsky, A. (1993) A yeast protein similar to bacterial two-component regulators. Science 262: 566-569.

Parkinson, J.S., and Kofoed, E.C. (1992) Communication modules in bacterial signaling proteins. Annu. Rev. Genet. 26: 71-112.

Pierson, B.K., and Castenholz, R.W. (1974) A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen. and sp. nov. Arch. Microbiol. 100: 5-24.

Ribeiro, A., Akkermans, A.D.L., van Kammen, A., Bisseling, T., and Pawlowski, K. (1995) A nodule-specific gene encoding a subtilisin-like protease is

expressed in early stages of actinorhizal nodule development. *Plant Cell* 7: 785-794.

Roberts, L.W., and Miller, A.R. (1981) Is ethylene involved in xylem differentiation? *Vistas in Plant Sciences*

Rodriguez, F.I., Esch, J.J., Hall, A.E., Binder, B.M., Schaller, G.E., Bleeker, A.B. (1999) A copper cofactor for the ethylene receptor ETR1 from *Arabidopsis*. *Science* 283: 996-998.

Roman, G., Lubarsky, B., Kieber, J.J., Rothenberg, M., and Ecker, J. (1995) Genetic analysis of ethylene signal transduction in *Arabidopsis thaliana*: five novel mutant loci integrated into a stress response pathway. *Genetics* 139: 1393-1409.

Sakai, H., Hua, J., Chen, Q.G., Chang, C., Medrano, L.J., Bleeker, A.B., and Meyerowitz, E.M. (1998) *ETR2* is an *ETR1*-like gene involved in ethylene signaling in *Arabidopsis*. *Proc. Nat. Acad. Sci. USA* 95: 5812-5817.

Salzberg, S.L., White, O., Peterson, J., and Eisen, J.A. (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science* 292: 1903-1906.

Schaller, A., and Ryan, C.A. (1995) Systemin - a polypeptide defense signal in plants. *BioEssays* 18: 27-33.

Schaller, G.E., and Bleeker, A.B. (1995) Ethylene binding sites generated in yeast expressing the *Arabidopsis* ETR1 gene. *Science* 270: 1809-1811.

Schaller, G.E., Ladd, A., Lanahan, M.B., Spanbauer, J.M., and Bleeker, A.B. (1995) the ethylene response mediator ETR1 from *Arabidopsis* forms a disulfide-linked dimer. *J. Biol. Chem.* 270: 12526-12530.

Shinde, U., and Inouye, M. (1994) The structural and functional organization of intramolecular chaperones: the N-terminal propeptides which mediate protein folding. *J. Biochem.* 115: 629-636.

Siezen, R. J., and Leunissen, J. A. M. (1997) Subtilases: the subtilisin-like serine proteases. *Prot. Sci.* 6: 501-523.

Takayama, S., and Sakagami, Y. (2002) Peptide signals in plants. *Curr. Opin. Plant Biol.* 5: 382-387.

Tanaka, H., Onouchi, H., Kondo, M., Hara-Nishimura, I., Nishimura, M., Machida, C., and Machida, Y. (2001) A subtilisin-like serine protease is required for epidermal surface formation in *Arabidopsis* embryos and juvenile plants. *Development* 128: 4681-4689.

Taylor, A. A., Horsch, A., Rzepczyk, A., Hasenkampf, C. A., and Riggs, C. D. (1997) Maturation and section of a serine proteinase is associated with events of late microsporogenesis. *Plant J.* 12: 1261-1271

Thacker, C., and Rose, A.M. (2000) A look at the *Caenorhabditis elegans* Kex2/subtilisin-like proprotein convertase family. *BioEssays* 22: 545-553.

Toure, B.B., Munzer, J.S., Basak, A., Benjannet, S., Rochemont, J., Lazure, C., Chretien, M., and Seidah, N. G. (2000) Biosynthesis and enzymatic characterization of human SKI-1/S1P and the processing of its inhibitory prosegment. *J. Biol. Chem.* 275: 2349-2358.

Whelan, S., Lio, P., and Goldman, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* 17: 262-272.

Wilson, I.B.H. (2002) Glycosylation of proteins in plants and invertebrates. *Curr. Opin. Struct. Biol.* 12: 569-577.

Winkel-Shirley, B. (2002) Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* 5: 218-223.

Xiong, J., Fischer, W.M., Inoue, K., Nakahara, M. and Bauer, C.E. (2000) Molecular evidence for the early evolution of photosynthesis. *Science* 289: 1724-1730.

Yamagata, H., Masuzawa, T., Nagaoka, Y., Ohnishi, T., and Iwasaki, T. (1994) Cucuminisin, a serine protease from melon fruits, shares structural homology with subtilisin and is generated from a large precursor. *J. Biol. Chem.* 269: 32725-32731.

Yamagata, H., Uesugi, M., Saka, Kazunori, K., Iwasaki, T., and Aizono, Y. (2000) Molecular cloning and characterization of a cDNA and a gene for subtilisin-like serine proteases from rice (*Oryza sativa* L.) and *Arabidopsis thaliana*. *Biosci. Biotechnol. Biochem.* 64: 1947-1957.

Zhao, C., Johnson, B.J., Kositup, B., and Beers, E.P. (2000) Exploiting secondary growth in *Arabidopsis*. Construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. *Plant Physiol.* 123: 1185-1196.

Appendix 1. Ethylene Receptor Protein Sequences

Protein	gi #	Description
ZeaETR2	10241927	<i>Zea mays</i>
StrawberryERS1	15131533	<i>Fragaria x ananassa</i>
VitisETR1	7547007	<i>Vitis vinifera</i>
TobaccoERS	7652766	<i>Nicotiana tabacum</i>
StrawberryEIN1	15131533	<i>Fragaria x ananassa</i>
RiceERS2	18252627	<i>Oryza sativa (indica)</i>
PyruscommunisETR1ii	18252319	
PyruscommunisETR1	18252351	
PyruscommunisERS1ii	18252335	
PyruscommunisERS1	18252345	
PrunusETR1	6841075	<i>Prunus persica</i>
PotatoERS	4210924	<i>Solanum tuberosum</i>
PhalaenopsisETR1	4650821	<i>Phalaenopsis sp.</i>
PetuniaETR1iii	17646117	
PetuniaETR1ii	17646115	
PetuniaETR1i	17646113	
PelargoniumETR2	5006606	
PelargoniumETR1	5006604	
PassionFruitERS2	19032291	<i>Passiflora edulis</i>
PapayaETR1	11935116	<i>Carica papaya</i>
OryzaERS?indica	15054446	
MangoETR1	7407123	<i>Mangifera indica</i>
LycopersiconETR6	19421763	<i>Lycopersicon esculentum</i>
LycopersiconETR2	2852993	<i>Lycopersicon esculentum</i>
LycopersiconETR1	2852991	<i>Lycopersicon esculentum</i>
DelphiniumERSii	18539224	<i>Delphinium</i>
DelphiniumERSi	18539222	<i>Delphinium</i>
CucumissativusETR2	6136818	<i>Cucumis sativus</i>
CucumissativusETR1	6136814	<i>Cucumis sativus</i>
CucumissativusERS	6136816	<i>Cucumis sativus</i>
CucumismeloETR1	11611626	
CucumismeloERS	2707334	<i>Cucumis melo</i>
BrassicaoleraceaETR2	18265377	
BananaEXX	6425101	<i>Musa acuminata</i>
AvocadoERS1	20386499	<i>Persea americana</i>
AppleERS	20135556	<i>Malus x domestica</i>
VignaradiataERS1	4138853	<i>Vigna radiata</i>
LycopersiconNVR	1122445	<i>ethylene receptor</i>
TobaccoETR1	2655061	<i>Nicotiana tabacum</i>
RumexpalustrisERS	2353690	<i>Rumex palustris</i>
PisumsativumERS	3123666	

PassionFruitETR1	4164159	<i>Passiflora edulis</i>
RiceERS	2281705	<i>Oryza sativa</i>
LycopersiconETR5	4877653	
LycopersiconETR4	4877651	
DianthusETR1	2541938	<i>Dianthus caryophyllus</i>
DianthusERS2	2662475	<i>Dianthus caryophyllus</i>
DianthusERS	2541968	<i>Dianthus caryophyllus</i>
CitrussinensisERS	4092526	<i>Citrus sinensis</i>
BrassicaoleraceaETR1	2896029	<i>Brassica oleracea</i>
BrassicaoleraceaERS	2896031	<i>Brassica oleracea</i>
ArabidopsisETR2	3687654	<i>Arabidopsis thaliana</i>
ArabidopsisETR1	902360	<i>ETR1 gene product</i>
ArabidopsisERS2	3687656	<i>Arabidopsis thaliana</i>
ArabidopsisERS	1046225	<i>ethylene response sensor</i>
ArabidopsisEIN4	4105633	<i>Arabidopsis thaliana</i>
AppleETR1	3411051	<i>Malus x domestica</i>
StrawberryETR1	15131529	<i>Fragaria x ananassa</i>

Appendix 3. Subtilisin-like Protein sequences used.

gi Number	Organism	Protein Name	Proposed Function	References
13430745	<i>Arabidopsis thaliana</i>	ARA12 (At5g67360)		
22331076		At3g14067		
18396192		At2g05920		
18423315		At5g51750		
18400322		At3g14240		
18418551		At4g34980		
18379241		SDD1 (At1g04110)	Regulation of stomatal pattern and density	Genes Dev. 14: 1119-31
18378894		At1g01900		
18425149		At5g67090		
16444943		ALE1 (At1g62340)	Epidermal surface formation	Development 128: 4681-9
18417514		At4g30020		
18398958		At2g19170		
18422451		At5g44530		
18415500		At4g20430		
18175991		At1g30600		
18424191		At5g59090		
20260319		At5g59120		
18424159		At5g58830		
18424157		At5g58820		
18424161		At5g58840		
18424155		At5g58810		
18424199		At5g59130		
18408293		At3g46840		
18408295		At3g46850		
18424193		At5g59100		
18414365		At4g15040		
18424207		At5g59190		
18411254		XSP1 (At4g00230)	xylem differentiation	Plant Physiol 123:1185
18414296		At5g03620		
18405184		At2g39850		
18394832		At1g20160		
18394830		At1g20150		
15238537		At5g59810		
22325457		AIR3 (At2g04160)	lateral-root formation	Plant Mol. Biol. 39: 273-287
15242454		At5g45640		

15242456	At5g45650	
18413349	At4g10530	
18413347	At4g10520	
18413345	At4g10510	
18413353	At4g10550	
18413351	At4g10540	
18398657	At1g32950	
18398655	At1g32940	
18398660	At1g32960	
15223351	At1g32970	
15223353	At1g32980	
15234556	At4g21640	
18415734	At4g21630	
18415740	At4g21650	
18408460	At1g66210	
18408462	At1g66220	
18415673	At4g21326	
18416719	At5g11940	
18415671	At4g21323	
18416730	At4g26330	
15241149	At5g19660	
9757901		
14488360	<i>Oryza sativa</i>	
20503056		
21741553		
9453867		
20146761		
22324424		
20160863		
20160864		
14209565		
6721520		
24414066		
20521377		
12328490		
21740609		
21740607		
21740608		
9957714	SP1	Biosci. Biotechnol. Biochem 64: 1947
20146283		
20160948		

20160950			
20160949			
22773236			
1771160	<i>Lycopersicon</i> <i>n</i> <i>esculentum</i>	SBT1	J.Biol.Chem. 275:5193
1771162		SBT2	
3183979		P69C1	Plant Mol.Biol. 39:749-760
7435660		P69E	
4200340		P69D	J.Biol.Chem. 274:2360-2365
4200334		P69A	J.Biol.Chem. 274:2360-2365
4200338		p69C	J.Biol.Chem. 274:2360-2365
4200336		P69B	J.Biol.Chem. 274:2360-2365
7435661		P69F	
6723683		P69F1	Plant Physiol 122:67-74
6723683		P69E1	Plant Physiol 122:67-74
3970733		SBT4C	Plant Mol. Biol. 39: 749-760
3687303		SBT4A	Plant Mol. Biol. 39: 749-760
3970757		SBT4E	Plant Mol. Biol. 39: 749-760
3970731		SBT4B	Plant Mol. Biol. 39: 749-760
3687301		SBT3	Plant Mol. Biol. 39: 749-760
6468325		TMP	Genome 44: 644-650
13325079	<i>Glycine max</i>	C1	Seed-reserves mobilization Plant Physiol. 114: 291
14091078		SSTP2	Physiol. Plant. 115: 585-597
7543934		SCS1	Planta 211: 484-492
1086249	<i>Alnus</i> <i>glutinosa</i>	Ag12	Actinorhizal nodule development Plant Cell 7: 785-794
431176	<i>Lilium</i> <i>longiflorum</i>	LIM9	Induced during meiosis DNA Res. 1: 15-26

807698	<i>Cucumis melo</i>	Cucumisin		J. Biol. Chem. 269: 32731-32731
7435682	<i>Picea abies</i>	AF70	antifreeze-like protein	
16080860	<i>Bacillus subtilis</i>	Bsu3805 (ipa-45r)		
16078384		Bsu1321 (ispA)		
22969863	<i>Chloroflexus aurantiacus</i>	Chlo0024		
21322744	<i>Streptomyces viridosporus</i>	DHP-A		Appl. Environ. Microbiol. 68: 2716-2725
1694627	<i>Streptomyces albogriseolus</i>			J. Bacteriol. 179: 430-438
20378355	<i>Drosophila Melanogaster</i>	dS1P		Dev. Cell 2: 229-238
21287806	<i>Anopheles gambiae</i>	agCP9458		
17369966	<i>Cricetulus griseus</i>	Site-1	Lipid metabolism regulation	Mol. Cell 2: 505-514
15079377	<i>Mus musculus</i>			
16758342	<i>Rattus norvegicus</i>	Site-1	Lipid metabolism regulation	Proc. Nat. Acad. Sci. U.S.A. 96: 1321-1326
4506775	<i>Homo Sapiens</i>	Site-1	Lipid metabolism regulation	Mol. Cell 2: 505-514

Appendix 5. Protein modelling templates

Protein	Templates (PDB codes)				
Air3	1vja	1vjb	1sbh	1vjc	1mpt
Ale1	1ah2	1jea	1st3	1svn	1c9jA
Ara12	1sbi	1sue	1duiA	1vjc	1sbh
At1g01900.batch1	1c31A	1scd	1be8	1scb	1sca
At1g01900.batch2	3vsb	1vsb	1selA	1selB	1avt
At1g04110.batch1	1s01	1sbh	1vjb	1vjc	1sbi
At1g04110.batch2	3prkE	1bjrE	1cnmA	1eggA	1ic6A
At1g20150	1duiA	1mpt	1sue	1spbS	1suaA
At1g20160	1duiA	1sue	1sbi	1suaA	1yjc
At1g30600	1ah2	1sibE	1jea	1st2	1s02
At1g32950.batch1	1sue	1suaA	1spbS	1sbi	1s02
At1g32950.batch2	1s02	1sud	1suc	1duiA	1sibE
At1g32960.batch1	1sue	1suaA	1sbi	1s02	1spbS
At1g32960.batch2	1s02	1spbS	3sicE	1sbnE	2st1
At1g32970.batch1	1sue	1suaA	1spbS	1duiA	1s01
At1g32970.batch2	1c9jA	1mpt	1ah2	1jea	1st3
At1g32980	1sue	1suaA	1spbS	1duiA	1sbi
At1g66210	1sue	1sbi	1suaA	1spbS	1sibE
At1g66220.batch1	1dbiA				
At1g66220.batch2	1sue	1sbi	1suaA	1s02	1spbS
At2g05920	1mpt	1duiA	1vjc	1vja	1vjb
At2g19170	1mpt	1svn	1jea	1ah2	1c9jA
At3g14067	1duiA	1sbi	1sue	1suaA	1spbS
At3g14240.batch1	1duiA	1vjc	1sbh	1vjb	1vja
At3g14240.batch2	1sbt	1svn	1ah2	2sbt	1tecE
At3g46840	1suaA	1spbS	1sue	1duiA	1c31A
At3g46850	1sbt	2sbt	1sbi	2sniE	1sue
At4g00230.batch1	1c31A	1bfu	1be6	1scd	1be8
At4g00230.batch2	1selA	1avt	3vsb	1av7	1vsb
At4g10510.batch1	1sue	1suaA	1s02	1sbi	1spbS
At4g10510.batch2	1suaA	1s02	1sbi	1spbS	1st2
At4g10520	1sue	1suaA	1sbi	1s02	1spbS
At4g10540	1sue	1suaA	1sbi	1s02	1spbS
At4g10550.batch1	1sue	1suaA	1s02	1sbi	1spbS
At4g10550.batch2	1s02	1sbi	1spbS	3sicE	5sicE
At4g15040	1c9nA	1scb	1scd	1scnE	1sca
At4g20430	1mpt	1st3			
At4g21323	1sbi	1s02	2sicE	2sbt	5sicE
At4g21630	1sbi	1sbt	2sbt	2sniE	1s02
At4g21640	1sbi	1sue	1meeA	1suaA	1sbh
At4g21650	1sue	1suaA	1sbi	1s02	1spbS

At4g26330	1scd	1scnE	1scb	2secE	1af4
At4g34980	1duiA	2sbt	1sbt	1sue	1yjb
At4g30020	1mpt	1jea	1ah2	1svn	1c9jA
At5g11940	1suaA	1spbS	1sue	1sbi	1s01
At5g19660.batch1	2sbt	1sbt	1yjb	1sbh	1yjc
At5g19660.batch2	1ic6A	1ht3A	1pekE		
At5g45640	1st3	1svn	1ah2	1c9jA	1jea
At5g45650	1sbi	1vja	2sniE	1yjb	1sbh
At5g51750	1yja	1yjc	1sbh	1yjb	1duiA
At5g58810	1duiA	1sue	1sbi	1spbS	1sbh
At5g58820	1yjc	1yja	1yjb	1duiA	1sbh
At5g58830	1duiA	1sbi	1sue	1yja	1spbS
At5g58840	1yja	1sbh	1yjc	1yjb	1duiA
At5g59090	1sbi	2sbt	1yjc	2sniE	1yjb
At5g59100	1sue	1suaA	1yjc	1yja	1sbh
At5g59120	1mpt	1jea	1st3	1svn	1ah2
At5g59130	1duiA	1sue	1yjc	1yja	1sbh
At5g59190	1duiA	1sbi	1sue	1spbS	1suaA
At5g67090	1mpt	1duiA	1c9jA	1st3	1jea